Functional linear mixed effects model for mode-of-action clustering in toxicity assessment

by

Tiantian Ma

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Statistics

Department of Mathematical and Statistical Sciences
University of Alberta

Abstract

Real Time Cell Analysis (RTCA) technology is used to monitor cellular changes continuously over the entire exposure period to chemicals. In RTCA system, chemicals with different concentrations are applied and time-dependent concentration response curves (TCRCs) are generated. In this thesis, we aim to study the mode of action (MOA) of tested chemicals by extracting important information from TCRCs and then do MOA clustering. In order to reduce the number of parameters to be estimated when fitting the data with limited sample size and high dimension, linear mixed effects models are applied by considering chemicals as random effects. The estimated and predicted coefficients from individual curves can be plugged into K-means and Self-organising maps to do clustering. Estimating curves using different functional bases corresponds to linear transformations of the data, and obtains information from various aspects of the curves. In this thesis, two different functional bases are used when fitting linear mixed models. The first model is based on functional principal components, which can stretch the data on a few directions that contain almost all the information. The two largest clusters, cluster 1 and cluster 10, can be separated with 88.24% accuracy rate on only two primary basis functions. According to the shape of the two functions and the coefficients distribution on them, we can depict the primary difference between clusters in terms of overall shape and local features. To detect the primary time intervals where the difference lies in, the other basis applied to mixed model is B-spline basis because different splines are dominant in different time intervals. The coefficients of spline basis perform well as input in both binary and multi-cluster clustering, with the clustering accuracy rate in the range of 81.82% to 86.54%. Those clustering results can be obtained by only using 1 to 2 primary directions in terms of time interval and concentration level, which is helpful to establish targeted experiments for further toxicants study.

Keywords: Time-dependent concentration response curves; Mode of action; Linear mixed effect model; Functional principal component analysis; B-spline; K-means; Self-organising map

Acknowledgement

First of all, I want to give my greatest thanks to Dr.Christina Anton and Dr.Adam Kashlak. They, as my supervisors, undoubtedly performed their works very well. Because of their unreserved help and patience, I can complete my studies. They also helped me plan my future development.

I sincerely thank the Mathematics and Statistics Department of the University of Alberta for giving me this opportunity to pursue academic progress, and I also thank you for your financial support. I also want to thank every lovely professor and classmate. Every academic sharing with them is unforgettable. In addition, a thank you to Yongqing Yang and Dan Richard, who gave me a lot of help at the beginning of my project.

I also would like to thank my parents. They encourage me to continue to pursue academic progress, and continue to praise me and give me confidence. My heartfelt thanks!

Contents

1 Introduction **1**

1.1 Introduction to functional data and their analysis 1

1.2 Introduction to TCRCs 2

1.3 Scope of this work 4

2 Statistical models and methods **7**

2.1 Functional Principal Component Analysis 7

2.1.1 PCA for multivariate data 7

2.1.2 FPCA for functional data 8

2.2 Functional linear mixed effect model 11

2.2.1 Linear mixed effect model for scalar data 11

2.2.2 Linear mixed effect model for functional data 13

2.3 Clustering analysis 15

2.3.1 K-means 15

2.3.2 Self-organising map 16

2.4 Spline 17

2.4.1 Cubic spline interpolation 17

2.4.2 B-spline 17

3 Apply FPCA-based FLMM to toxicity data **18**

3.1 Model and model assumptions 19

3.2 Estimation results from R 22

3.3 Reconstruct random effects 27

3.4 Binary clustering 31

3.4.1 Visual clustering by score plots 31

3.4.2 Test difference by Manova 32
3.4.3 Clustering by k-means 34
3.4.4 Clustering by SOM 36

4 Apply FLMM with B-spline basis to toxicity data 39

4.1 Model and model assumptions 40
4.2 Estimation results from R 41
4.3 Reconstruct the TCRCs 43
4.4 Binary clustering 43
4.4.1 Clustering by k-means 44
4.4.2 Clustering by SOM 48
4.5 Multi-cluster clustering 50
4.5.1 Clustering by k-means 50
4.5.2 Clustering by SOM 52

5 Discussion and summary 54

List of Figures

1.1 An example of the original data for two chemicals 4

3.1 TCRCs from two different clusters 19

3.2 Point-wise variation curves per concentration 20

3.3 Point-wise mean curves for chemicals in cluster1 per concentration (upper left), Point-wise means for chemicals in cluster10 per concentration (upper right), Point-wise mean curves for chemicals in cluster10 per concentration (upper right), Point-wise variance curves for chemicals in cluster1 per concentration (bottom left), Point-wise variance curves for chemicals in cluster10 per concentration (bottom right) . . . 22

3.4 FPCs for cluster1 chemical effects (left) and the cluster10 chemical effects (right): the first 3 FPCs of con1-specific effects (the first row), he first 3 FPCs of con2-specific effects (the second row), the first 3 FPCs of con3-specific effects (the third row), the first 3 FPCs of smooth error term (the fourth row) 23

3.5 FPCs for all chemical effects: the first 3 FPCs of con1-specific effects (the first row), he first 3 FPCs of con2-specific effects (the second row), the first 3 FPCs of con3-specific effects (the third row), the first 3 FPCs of smooth error term (the fourth row) 24

3.6 Concentration-specific means (black) plus (blue) and minus (red) chemical effects estimated by the first three components of concentration 1 (upper 3 figures), concentration 2 (middle 3 figures), concentration 3 (bottom 3 figures). The respective proportion of variability induced by chemical effects within concentration is given in brackets 26

3.7 Conc1-specific chemical effects (cluster1) 27

3.8 Conc1-specific chemical effects (cluster10) 28

3.9 Real data VS estimated data of 4 chemicals (cluster1) 29

3.10 real data VS estimated data of 4 chemicals (cluster10) 29
3.11 Chemical-20 in concentration 3 (the first row) and chemical-26 in concentration3 (the second row). For each row, the left figure is real data; the middle is estimated data by first 2 FPC's; the right is estimated data by first 3 FPC's 30
3.12 Score plots for 3 concentrations. 31
3.13 Chi-square Q-Q plot 33
3.14 An example of clustering result by SOM 38

4.1 Each black curve corresponds to the average of a chemical TCRCs in a specific concentration level. The red curve is the average of black curves 40
4.2 Raw TCRCs VS predict TCRCs 43
4.3 2-dimensional plot of coefficients (cluster 1 VS cluster 10) 45
4.4 2-dimensional plot of coefficients (cluster 1 VS cluster 11) 46
4.5 1-dimensional plot of coefficients (cluster 10 VS cluster 11) 48
4.6 2-dimensional plot of coefficients (cluster 1 VS cluster 10 VS cluster 11) 51
4.7 An example of clustering result by SOM 53

List of Tables

3.1 Variance decomposition 25
3.2 Manova 33
3.3 Clustering result by k-means 35
3.4 Clustering results by SOM with different parameters 37
3.5 Accuracy rate summary 37

4.1 Summary of linear mixed effect model (cluster 1 VS cluster 10) 42
4.2 Best clustering result (cluster 1 VS cluster 10) 44
4.3 Different inputs in K-means and the corresponding results (cluster 1 VS cluster 10) . 44
4.4 Best clustering result (cluster 1 VS cluster 11) 46
4.5 Different inputs in K-means and the corresponding results (cluster 1 VS cluster 11) . 46
4.6 Best clustering result (cluster 10 VS cluster 11) 47
4.7 Different inputs in K-means and the corresponding results (cluster 10 VS cluster 11) 47
4.8 Accuracy rate summary (cluster 1 VS cluster 11) 49
4.9 Best clustering result by K-means (cluster 1 VS cluster 10 VS cluster 11) 50
4.10 Different input in Kmeans (cluster 1 VS cluster 10 VS cluster 11) 50
4.11 Accuracy rate summary (cluster 1 VS cluster 10 VS cluster 11) 52
4.12 Best clustering result by SOM (cluster 1 VS cluster 10 VS cluster 11) 53

Chapter 1

Introduction

1.1 Introduction to functional data and their analysis

Technological progress makes it possible for scientists in various fields to collect and store a growing amount of functional data. These data have a functional nature because they can, at least theoretically, be observed in arbitrarily fine resolution. Most commonly, these data are real-valued one-dimensional curves observed over time. They can also be collected on higher-dimensional domains such as surfaces or shapes. Giving the availability of functional data, a new branch of statistics motivated by the desire to explore the potential of these data called functional data analysis (FDA [18]) stepped onto stage.

In practice, the data at hand consist of vectors of discrete observations instead of continuous functions, which look the same to multivariate data. However, the key difference between them is that functional data are structured objects with a natural ordering in their dimensions rather than a collection of single data points. FDA accounts for the natural ordering by treating functional data as realizations of a stochastic process, with smoothness assumed to reflect the similarity of adjacent values [18]. Applications of FDA are numerous and come from diverse fields, including acoustic research in speech science, spectroscopy study in chemistry or medicine, climate and neuroimaging data study, etc. All this raises the need to extend scalar and multivariate practical methods to functional data analysis. For example, regression modelling, classification and clustering approaches and dimension reduction methods, etc.

The other typical application of FDA is toxicity assessment studied in this work. As we all know, the increasing number of chemical compounds in the environment may impose hazard effects on human health. Some chemicals may cause toxic effects on cells such as apoptosis and necrosis [16],

while others may induce uncontrolled cellular proliferation [1]. Therefore, the important indicator in assessing toxicity is the cellular change, where time plays a key role in the dynamic process of interaction between chemicals and cells. In order to study the property of chemicals, the bio-activity indices of cells need to be observed and recorded along the time, which is essentially real-valued functional data. The study based on this data is in the scope of FDA.

1.2 Introduction to TCRCs

In the past, data used for assessing toxicity was observed and collected by applying toxic chemicals on animals. Such experiments implemented on living organism are called in vivo assays [6], which require a large number of samples to be observed and may cause biological contamination. Recently, Cell-based in vitro assays [25] are universally applied. Compared with traditional in vivo assays, in vitro alternatives are easier to carry out and less time consuming. Thus, there is an increase need for developing effective methodologies to analyze large amount of data from in vitro assays.

The data set in this study were generated from one such in vitro assay -xCELLigence Real-Time Cell Analysis system (RTCA) [13] developed by the ACEA Biosciences Inc. (San Diego, USA). This system utilizes 384 well electronic plates (E-Plates 384), on the bottom of which the electric current is impeded by cells attached to electrodes. The impedance data is recorded as direct measurement of cellular status in real time including cell number, cell morphology and cell adhesion [20]. The data is converted from impedance to Cell Index (CI) by the following formula [7, 12]

$$CI = \max_{k=1,\ldots,K} \left[\frac{R_{cell}(f_k)}{R_b(f_k)} - 1 \right] \tag{1.1}$$

where $R_{cell}(f_k)$ and $R_b(f_k)$ are the electrode impedance with and without cell in the well, and k is the discrete time point.

Our work is based on the data from the cytotoxicity profiling project carried by Alberta Center for Toxicity. They carried out assays by the RTCA system, in which the growth of human hepato carcinoma cells line (HepG2) were tested with 65 chemicals [23] added in basal media. Each testing chemical obtained was at least 95% purity and then diluted into 11 concentrations for single usage. In order to give cells a suitable growth environment, the assays simulated the human environment and were carried out in 37°C tissue culture hood with 95% humidity and 5% CO2. The assays were

implemented in the E-plate 384 and started by seeding the HepG2 cells in the wells. 24 hours later, when the cells attached to the bottom and adapted to the culture environment, 11 concentrations of each testing chemical were applied. The RTCA system continuously monitored the impedance signal in the wells for at least 76 hours, such that time-dependent concentration response curves (TCRCs) for each test chemical were generated by converting the impedance data to CI [23].

TCRCs provide useful information about the mechanism of interaction between cells and testing chemicals. This mechanism is referred to as mode of action (MOA) describing series of physiological events, such as cell population, cell morphology and cellular functions [23]. Since the TCRCs don't provide much information of MOA before chemicals are added, they are truncated such that only the data after treatment were kept. Different TCRCs may have different CI values at 24 hours. What we focus on is the cellular response to testing chemicals, thus CI differences from chemical adding and growth were minimised by using Normalized Cell Index (NCI), which is given by Zhang et al [29] as formula(2).

$$NCI[k] = \frac{CI[k]}{CI[0]}, k = 1, 2, ..., K \tag{1.2}$$

where k refers to different time points after testing chemical addition, and k=0 refers to the time point right before treatment. Using NCI also make sure all TCRCs start from almost the same level around 1.

An example of the TCRCs after truncation for two chemicals is given in Figure 1.

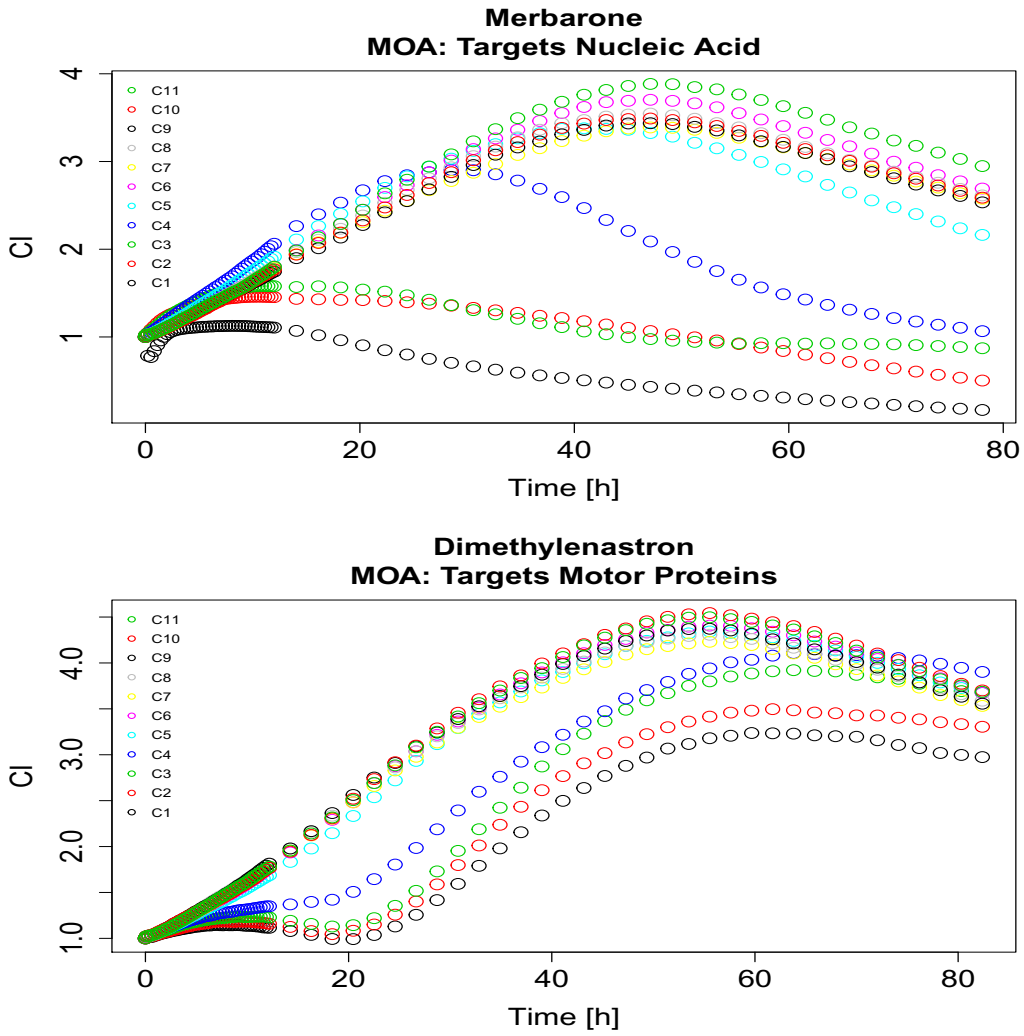


Figure 1.1: An example of the original data for two chemicals

It is noticed that time grids are not uniform in raw TCRCs. Cubic splines were fit by Zhang et al. [29] to interpolate the non-uniform data into uniform grids . From each spline 161 evenly spaced points (in time) were sampled to form the spline data. The spline data set is used for further analysis in this study.

1.3 Scope of this work

In this study, we focus on MOA clustering for the 65 chemical compounds given by Alberta Center for Toxicology. According to the MOA, the 65 chemicals were divided into 10 clusters and the list of ten-cluster MOA classification was also provided. For example, in figure 1, "Merbarone" and "Dimetheylenastron" have different MOA clusters, the former targeting nucleic acid, while the

later targeting motor proteins. By comparing the two chemicals in figure 1, the TCRCs show much similarity of some concentrations, while the TCRCs trend of concentration1-4 are very different. It is also noticed that even if the overall trend for concentration1-4 TCRCs of the two chemicals are quite different, the local curves at the very beginning (0 - 20 hours) have almost the same pattern. Thus, we can see that some concentrations and time intervals play the import role in clustering while others contribute a little to our study.

From the information included in the TCRCs, some indices such as LC_{50}, KC_{50} and AUC_{50} were extracted and used to describe the time and concentration dependent cellular activity. LC_{50} represents the concentration that can kill 50% of cells [27]. KC_{50} uses an exponential model to calculate the LC_{50} value [27, 28, 14]. AUC_{50} is used to represent the area under normalized TCRCs, which is an index of toxicity evaluation [13]. These indices were also used in statistical methods for clustering and classification. But these indices only provide partial information of TCRCs and some significant features may not be uncovered. In order to extract more information and reduce the input data, wavelet transform of TCRCs was applied to MOA classification by using the wavelet coefficients as input into machine learning methods such as artificial neurol network (ANN) and support vector machine (SVM) [29]. The similar objective was also achieved by using the first principal component scores of TCRCs in model-based classification approach [24]. I also achieve this goal by applying a linear mixed model based on functional principal component analysis to TCRCs. The advantage of using a mixed effect model is that information can be separately extracted from different sources of variability instead of pooled variance. Thus, we can focus on the information we are interested in and ignore the redundant information induced by replication and measurement error. Moreover, the number of functional principal components (FPCs) for clustering was decided by variance. The reconstructed TCRCs using truncated FPCs perform well in capturing main features of the original TCRCs, which ensures that valid information is retained and reduce the data dimension at the same time. A similar mixed effect model was also applied in tissue spectroscopy study [2], where the correlation of the data comes from the process of data collection. By separating the variance into different parts, the contribution of the tissue type, which is the factor they are interested in, is extracted. Another similar study was motivated by the data of coronary sinus potassium concentrations measured on 4 treatment groups of dogs [22]. In this study, the variance induced from dogs are accounted for by regarding dog as random effect in the mixed model.

According to the demonstration above, the main objectives of my work are twofold. First,

develop a model capable of reconstructing the TCRCs, and based on the model, we are able to extract the effects of tested chemicals on cell population. These chemical effects can be used to perform clustering. The clustering result is then validated with the known MOA clusters. Second, capture and utilize the important features, for example TCRCs of interesting concentrations and time intervals, to distinguish different clusters, such that the clustering result is as good as that using the whole TCRCs.

An important contribution of my work is choosing the hierarchical functional linear mixed effect model to fit the data, such that the variability of the data is broken down to different sources induced by concentration levels, individual chemicals, replications and measurement error. By this decomposition, we can focus on chemical effects and take advantage of information on the variability of chemicals. In the first linear mixed effect model, functional principal component analysis is applied to reduce dimension by constructing a new coordinate system in a way that the largest variance is determined by the first principal component, the second largest variance on the second principal component, and so on. It turns out that only using the first 3 components to reconstruct the data can capture most features of the original TCRCs. The clustering results are also satisfactory by using the scores of the first 3 components as input to k-means algorithm or self-organized map. Since clustering is performed concentration-by-concentration, those concentrations where the significant differences of TCRCs exist can be located. In the second functional linear mixed model, I specify B spline basis with 4 spline functions as a new coordinate system. The reconstructed curves and clustering result are also satisfactory by using the coefficients of the 4 basis functions. Most importantly, with the characteristic of spline basis where different splines are dominant in different time intervals, we can find the time intervals where significant differences of TCRCs exist.

The remainder of the thesis is organised as follows. In chapter two, statistical models and methods are introduced, including linear mixed effect model and principal component analysis, as well as their functional counterparts, clustering analysis and two clustering algorithms, spline interpolation and B-spline basis. The next chapter focuses on applying linear mixed model based on functional principal component analysis (FPC-based FLMM [2]) to the data set. The model is validated by reconstructing the TCRCs. Different clustering methods such as k-means and self-organised map are applied to the FPC scores. In the fourth chapter, functional linear mixed effect model with B spline basis is used. The coefficients on the splines are chosen as input to perform clustering. In the last chapter, main results are summarized and discussed.

Chapter 2

Statistical models and methods

This chapter gives the introduction of statistical models and methods used in the thesis. Firstly, functional principal component analysis is introduced, followed by functional linear mixed effects models, including model assumptions and the outline of the estimation and prediction of the parameters. Clustering analysis and clustering algorithms such as K-means and self-organising map are explained. At last, I introduce spline interpolation and B-spline basis as preliminary knowledge for data processing and model fitting.

2.1 Functional Principal Component Analysis

FPCA is one of the most fundamental concepts of Functional Data Analysis(FDA) and it is a powerful tool for dimension reduction in terms of two aspects: FPC's are coordinates maximizing variability and an optimal orthonormal basis [5]. Coordinates maximizing variability describe FPC's as a projection of data to lower-dimensional space with maximizing retained variance. The basis resulting from FPCA is optimal because the expansion of each curve in terms of FPC's and corresponding scores approximates the original curve as closely as possible [17]. In order to illustrate how to derive FPC's and the properties of FPCA, we start from multivariate PCA, which is the counterpart of FPCA for multivariate data.

2.1.1 PCA for multivariate data

Before giving the procedure of PC's derivation, some notations are specified. Assume $\mathbf{x} \in \mathbf{R^p}$ to be a random vector with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. $\mathbf{X}$ is an $n \times p$ matrix with each row a sample from $\mathbf{X}$ and columns of $\mathbf{X}$ are centered. $\mathbf{S}$ is the sample variance-covariance

matrix defined by $\mathbf{S} = \frac{1}{n-1}\sum(\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})^{\mathbf{T}}$. $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x^T y}$ is the inner product of vector $\mathbf{x}$ and $\mathbf{y}$. $\|\mathbf{x}\| = \langle \mathbf{x}, \mathbf{x} \rangle^{\mathbf{1/2}}$ is the norm of $\mathbf{x}$.

The underlying idea of PCA is to find some orthonormal vectors (directions) $\xi_{\mathbf{k}}, k = 1, ..., p$ composing a basis, on which samples are projected and the variance on each of the direction is maximized. According to this principle, PC's can be derived as follows:

Step 1. Find a $\xi_{\mathbf{1}}$ such that projections $f_{i1} = \langle \xi, \mathbf{x_i} \rangle = \xi^{\mathbf{T}}\mathbf{x_i}, i = 1, ..., n$ on it have the maximum variance, i.e.

$$\begin{aligned} \xi &= \underset{\|\xi\|=1}{\operatorname{argmax}} \frac{1}{n-1}\sum_{i=1}^{n} f_{i1}^2 \\ &= \underset{\|\xi\|=1}{\operatorname{argmax}} \frac{1}{n-1}\sum_{i=1}^{n} (\xi^{\mathbf{T}}\mathbf{x_i})^{\mathbf{2}} = \underset{\|\xi\|=1}{\operatorname{argmax}} \frac{1}{n-1}\xi^{\mathbf{T}}\mathbf{X^T X}\xi \\ &= \underset{\|\xi\|=\mathbf{1}}{\operatorname{argmax}} \xi^{\mathbf{T}}\mathbf{S}\xi = \underset{\|\xi\|=\mathbf{1}}{\operatorname{argmax}} \langle \mathbf{S^T}\xi, \xi \rangle \end{aligned} \tag{2.1}$$

The optimization problem above is equivalent to finding a unit length vector ξ such that $\xi^{\mathbf{T}}\mathbf{S}\xi$ is maximal. According to the spectral decomposition[5], the maximum of $\xi^{\mathbf{T}}\mathbf{S}\xi$ is given by λ_1, the largest eigenvalue of $\mathbf{S}$ and $\xi_{\mathbf{1}} = \mathbf{u_1}$, the eigenvector of $\mathbf{S}$ corresponding to λ_1.

Step 2. For the subsequent PCs $\xi_{\mathbf{2}}, ...\xi_{\mathbf{p}}$, we follow the same principle of finding $\xi_{\mathbf{1}}$ and additional constraint(s) $\langle \xi_{\mathbf{k}}, \xi_{\mathbf{i}} \rangle = 0, 0 < i < k$, which guarantees $\xi_{\mathbf{k}}$ we are finding is orthogonal to the previous components. Similar to the derivation method of $\xi_{\mathbf{1}}$ component, $\xi_{\mathbf{k}}$ is given by the eigenvector of $\mathbf{S}$ corresponding to λ_k, the k^{th} largest eigenvalue.

We notice that deriving PC's is essentially a procedure to solve eigen problem of sample variance-covariance matrix $\mathbf{S}$. Analogously, if the counterpart of $\mathbf{S}$ for functional data is defined, the similar eigen analysis would be conducted to find FPC's.

2.1.2 FPCA for functional data

In the view of Horvath and Kokoszka(2012) [5], some functional data falls into the "large p small n" setting in the sense that every data object in such setting is measured by a large number of scalar values while the sample size n is much smaller than the number of measurements. Then,

how to use only m coefficients in a standard space to represent a large number of measurements per sample becomes an interesting topic, which can also make it feasible to apply multivariate analysis on the data.

Definitions in functional setting

Before carring over the idea of PCA to FPCA, some definitions based on Horvath and Kokoszka(2012) [5] and Jolliffe(2016) [9] for functional data are introduced.

Let $X(t) \in L^2(\mathcal{T})$ be a square integrable random process. For simplicity, assume $X(t)$ is centered. i.e. $\mu(t) = E[X(t)] = 0$. Other important definitions are as following:

$$\begin{aligned} &K(t,t^{'}) = E[X(t)X(t^{'})] \quad \text{(auto-covariance function)} \\ &C(t) = [Ky](t) = \int_{\mathcal{T}} K(t,t^{'})y(t^{'})dt^{'}, \quad y \in L^2(\mathcal{T}) \quad \text{(covariance operator)} \end{aligned} \tag{2.2}$$

The covariance operator is essentially a mapping that maps $y(t)$ from L^2 to L^2. Like the role of variance-covariance matrix Σ in PCA, $C(t)$ is the heart of FPCA. To have a better understanding of $C(t)$, let's rewrite Σ as a mapping $\Sigma : R^p \to R^p$ like Cederbaum and Jona [2] did in 2017:

$$[\mathbf{\Sigma y}]_i = \sum_{j=1}^{p} (\Sigma_{ij} y_j), \quad \mathbf{y} \in \mathbf{R^p}$$

We notice that the integration for the covariance operator is replaced by a summation, and augment t is replaced by index j. Thus, covariance operator in functional setting can be regarded as an analogue to variance-covariance matrix in mulitivariate setting.

Assume $x_i, \quad i = 1, ..., n$ are samples from $X(t)$. As needed in FPCA, samples are centered. Then the sample mean is $\hat{\mu}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t) = 0$. Sample counterparts for $K(t,t^{'})$ is defined as:

$$\hat{K}(t,t^{'}) = \frac{1}{n-1}\sum_{i=1}^{n}[x_i(t)x_i(t^{'})] \tag{2.3}$$

Derivation for FPC's

Similar to the idea of PCA, the aim of FPCA is to find some orthonormal functions (directions) $\xi_k(t), k = 1, ..., p$ composing a basis, on which samples are projected and the variance on each of the direction is maximized. According to this principle, FPC's can be derived in the following steps:

Step 1. Find a function $\xi_1(t)$ such that projections $f_{i1} = \langle \xi, x_i \rangle = \int_{\mathcal{T}} \xi_1(t) x_i(t) dt, i = 1, ..., n$ on it have the maximum variance, i.e.

$$\begin{aligned} \xi_1(t) &= \underset{\|\xi(t)\|=1}{\operatorname{argmax}} \frac{1}{n-1} \sum_{i=1}^{n} f_{i1}^2 \\ &= \underset{\|\xi(t)\|=1}{\operatorname{argmax}} \frac{1}{n-1} \sum_{i=1}^{n} \Big(\int_{\mathcal{T}} \xi(t) x_i(t) dt \Big)^2 \end{aligned} \tag{2.4}$$

By expanding the square term and changing the order of integral and summation, the above equation can be expressed as

$$\begin{aligned} \xi_1(t) &= \underset{\|\xi(t)\|=1}{\operatorname{argmax}} \frac{1}{n-1} \sum_{i=1}^{n} \int_{\mathcal{T}} x_i(t) \Big(\int_{\mathcal{T}} \xi(t^{'}) x_i(t^{'}) dt^{'} \Big) \xi(t) dt \\ &= \underset{\|\xi(t)\|=1}{\operatorname{argmax}} \int_{\mathcal{T}} \int_{\mathcal{T}} \Big(\frac{1}{n-1} \sum_{i=1}^{n} x_i(t) x_i(t^{'}) \Big) \xi(t^{'}) dt^{'} \xi(t) dt \\ &= \underset{\|\xi(t)\|=1}{\operatorname{argmax}} \int_{\mathcal{T}} \Big(\int_{\mathcal{T}} \hat{K}(t, t^{'}) \xi(t^{'}) dt^{'} \Big) \xi(t) dt \\ &= \underset{\|\xi(t)\|=1}{\operatorname{argmax}} \int_{\mathcal{T}} [\hat{K}\xi](t) \xi(t) dt \\ &= \underset{\|\xi(t)\|=1}{\operatorname{argmax}} \langle \hat{K}\xi, \xi \rangle \end{aligned} \tag{2.5}$$

It was proved that $K(t, t^{'})$ and its sample counterpart $\hat{K}(t, t^{'})$ are both symmetric, positive definite Trace class operators [5]. Then, the optimal problem (2.5) is essentially an eigen problem of $\hat{K}$. The supremum is reached if $\xi_1 = \nu_1$ and the maximum is λ_1, where λ_1 is the largest eigenvalue of $\hat{K}$ and ν_1 is the corresponding eigen function defined by $\hat{K}\nu_j = \lambda_j \nu_j$ (see, Horvath and Kokoszka, 2012, Chapter2, Chapter3[5])

Step 2. For the subsequent FPCs $\xi_2(t), ..., \xi_p(t)$, we follow the same principle of finding $\xi_1(t)$ with additional constraint(s) $\langle \xi_k(t), \xi_i(t) \rangle = \int_{\mathcal{T}} \xi_k(t) \xi_i(t) dt = 0, 0 < i < k$, which guarantees the $\xi_k(t)$ is orthogonal to the previous FPCs. Similar to the derivation method of $\xi_1(t)$, $\xi_k(t)$ is given by the eigenvector of $\hat{K}$ corresponding to λ_k, the k^{th} largest eigenvalue.

Another way to look at FPCA is based on Karhunen-Loeve expasion (KL expansion) [5]. KL expasion represents a continous stochasitc process as a linear combination of orthogonal functions. For the zero mean random process $X(t)$, the K-L expansion is given by

$$X(t) = \sum_{k=1}^{\infty} f_k \xi_k(t)$$

where ξ_k are orthonormal functions and $f_k = \langle \xi_k, x \rangle$ are uncorrelated zero mean random basis weights with variance λ_k.

2.2 Functional linear mixed effect model

Mixed effect models extend the predictor $X\beta$ of regression models by incorporating random effects in addition to fixed effects β [3]. Unlike regression models, which are based on the assumption that the data are independent and identically distributed, the introduction of random effects in mixed effect model makes it possible to capture the correlation of the data on the basis of the common regression characteristics in the population. Therefore, mixed effect models provide flexibility in analysis of data with multiple sources of variation, such as repeated measures, clustered or longitudinal data, or data with special structures. Random effects in the model can improve the performance of regression by accounting for the between- and within-subject variability of responses.

2.2.1 Linear mixed effect model for scalar data

In its general form, a scalar linear mixed model can be defined as

$$\mathbf{y} = X\beta + Zu + \epsilon \tag{2.6}$$

where $\mathbf{y} = (y_1, ..., y_n)^T$ is a vector of n observable random response variables, X and Z are known $n \times p$ and $n \times q$ design matrices, corresponding to the $p \times 1$ and $q \times 1$ vectors β and u of fixed and random effects, respectively.

A related situation to give some intuition of the model is the analysis of clustered data, i.e., when data are observed from subjects by subsampling from clusters or groups. For example, patients, students, or clients selected from hospitals, schools or firms, which are regarded as clusters. According to the model, β represents the population effects and $X\beta$ can be thought as population mean. While u is a vector of cluster-specific random effects, and Zu are deviations of each cluster

from the population mean. If equation (2.6) is a standard linear model, where u is assumed to be fixed, it is impractical in this case that the number of groups is relatively large, because the model would under perform when the number of parameters to be estimated becomes quite large relative to the sample size. Another advantage of mixed models is that the correlations induced by repeated observations from clusters are taken into account during estimation[3].

Model assumption

Since the population mean is represented by the fixed effects, u and ϵ are assumed to be independent and both have zero mean. A common assumption is that u and ϵ follow the multivariate Gaussian distribution

$$\begin{pmatrix} u \\ \epsilon \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} K & 0 \\ 0 & \Sigma \end{pmatrix} \right]$$

Estimation and prediction

In the package denseFLMM [2], the package I use to fit the model in next chapter, the way of estimation and prediction is maximizing the joint log-likelihood of y and u with respect to β and u, which is given by

$$l(\beta, u) = -\frac{1}{2}(\mathbf{y} - X\beta - Zu)^T \Sigma^{-1}(\mathbf{y} - X\beta - Zu) - \frac{1}{2}u^T K^{-1} u$$

Maximizing the joint likelihood can be converted to minimizing the penalized least square criterion where $(\mathbf{y} - X\beta - Zu)^T \Sigma^{-1}(\mathbf{y} - X\beta - Zu)$ is the weighted least squares criterion and $u^T K^{-1} u$ is the penalty term. Thus, the optimization of log-likelihood results in the following weighted least squares for the fixed effects

$$\begin{aligned} \widehat{\beta} &= (X^T V^{-1} x)^{-1} X^T V^{-1} \mathbf{y} \\ \widehat{u} &= K Z^T V^{-1} (\mathbf{y} - X\beta) \end{aligned} \tag{2.7}$$

where $V = ZKZ^T + \Sigma$. The above equations in terms of $\widehat{\beta}$ and $\widehat{u}$ are not closed-form expressions since we still don't know K and V. Let ν denote the vector of all variance parameters in the matrices K and Σ, thus in V. The estimation of variance parameters ν is commonly done using

maximum likelihood and the log-likelihood is given by

$$l(\beta,\nu) = -\frac{1}{2}\bigg(log|V(\nu)| + (\mathbf{y} - X\beta)^T V(\nu)^{-1}(\mathbf{y} - X\beta)\bigg) \tag{2.8}$$

In order to estimate ν, we rewrite the log-likelihood as the profile log-likelihood for ν, which is defined as $l_P(\nu) = max_\beta l(\beta,\nu)$. The ML estimate of ν is obtained by maximizing the profile log-likelihood $l_P(\nu)$ with respect to ν by replacing β with $\widehat{\beta}(\nu)$ as defined in (2.7). Equation (2.8) is not a closed-form expression either because β is unknown. Therefore, the maximization of $l_P(\nu)$ should be obtained by iteratively plug $\widehat{\nu}$ to equation (2.7) and $\widehat{\beta}$ to equation (2.8).

2.2.2 Linear mixed effect model for functional data

Ma and Zhong (2008)[11] consider what can be called a functional nonparametric mixed effect model of the form

$$Y_i(t) = \mu(t,\mathbf{x}_i) + \mathbf{z}_i^T\mathbf{U}(t) + \epsilon_i(t)), \qquad i = 1,...,n,\ t \in \mathcal{T} \tag{2.9}$$

where the population mean $\mu(t,\mathbf{x}_i)$ is assumed to be a smooth mean function dependent on scalar and/or functional covariates $\mathbf{x}_i$. $\mathbf{z}_i$ is a covariate vector. $\mathbf{U}(t)$ denotes a vector of functional random effects, which is a vector-valued zero mean, square integrable random process on $\mathcal{T}$. $\epsilon_i(t)$ represents white noise uncorrelated along $\mathcal{T}$. They are i.i.d. mean zero random variables with variance σ^2 for all i and t_{ij}. $\mathbf{U}(t)$ is independent of $\epsilon_i(t')$ for all i, t, t'. Usually, the functional random effects (FREs) include a smooth error term which is a functional random intercept (FRI) with a special structure that captures deviations from the mean function which are correlated along $\mathcal{T}$ [2].

Model (2.9) is a "piecewise" model, whose vector of fREs $\mathbf{U}(t)$ is divided into G independent blocks, one for each grouping factor, i.e., $\mathbf{U}(t) = [\mathbf{U}_1(t)^T,...,\mathbf{U}_G(t)^T]^T$. For each factor, different levels are represented by L_{U_g} independent copies $\mathbf{U}_{gl}(t), l = 1,...,L_{U_g}$. Each of independent copies consists of P_{U_g} FREs, yielding g^{th} block $\mathbf{U}_g = [U_{g11},...,U_{g1P_{U_g}},.....,U_{gL_{U_g}1},...,U_{gL_{U_g}P_{U_g}}]$. Thus, the total number of FREs in the model is $\sum_{g=1}^{G} L_{U_g}P_{U_g}$. For example, there are four patients from two hospitals. We regard hospital and patient as two grouping factors and the model includes a FRI for each hospital and a correlated FRI and FRS (functional random slope) for each patient. In this case, $G = 2$ (number of grouping factors), $L_{U_1} = 2$ (number of hospitals) and $L_{U_2} = 4$ (number of patients). The number of random effects associated with grouping factor g is P_{U_g}, $g = 1,2$. Thus, $P_{U_1} = 1$ (an FRI for each hospital) and $P_{U_2} = 2$ (an FRI and an FRS for each patient). The

number of FREs in this case is 10.

Model assumption

It is assumed that FREs in diffenrent blocks are independent and copies within each block are also independent. Only FREs of the same copy are correlated. In the above example of patients within hospitals, the FREs of hospitals are independent of the FREs of patients. Moreover, the FREs of different hospitals are independent and the FREs of different patients are also independent. Only the FREs of the same patient are correlated. Accordingly, the covariance of $\mathbf{U}(t)$ is a diagonal block matrix as following form

$$C_U(t,t') = diag\left[\underbrace{C_{U_1}(t,t'),...,C_{U_1}(t,t')}_{L_{U_1}\text{copies}},.....,\underbrace{C_{U_G}(t,t'),...,C_{U_G}(t,t')}_{L_{U_G}\text{copies}}\right] \tag{2.10}$$

where $C_{U_g}(t,t') = Cov\left[\mathbf{U}_{gl}(t),\mathbf{U}_{gl}(t')\right]$ is a $P_{U_g} \times P_{U_g}$ covariance matrix for $l = 1,...,L_{U_g}$.

Estimation and prediction

Estimation and prediction in the package denseFLMM are conducted by expanding FREs in functional principal component bases. Since the process of estimation involves matrix vectorization and Kronecker product of matrices $\otimes$, for ease of illustration, I only introduce the rough idea of each step in estimation process.

Step 1. Estimate the mean

If the mean $\mu(t)$ only depends on a discrete variable k, i.e., is a group-specific mean function $\mu_k(t)$ as the data set in my thesis, we can estimate it by simply averaging curves $Y_i(t)$ point-wise within each group k.

Step 2. Estimate the covariance structure

The covariance structure can be represent as $E\big[(\mathbf{Y}-\mu)\otimes(\mathbf{Y}-\mu)\big]$, where $\mathbf{Y}$ is matrix format of the response. The covariance can be decomposed as a linear combination of block covariance matrices C_{U_g} in (2.10), with coefficients being piece-wise covariate design matrices consisting of $\mathbf{z}_i$. Then, based on the equivalence of two representations of covariance structure, the covariance matrices $\widehat{C}_{Ug}, g = 1,...,G$ can be estimated based on least square.

Step 3. Estimate the eigenfunctions and eigenvalues

Estimation of eigenfunctions and eigenvalues are obtained by using spectral decomposition of the covariance $\widehat{C}_{Ug}$. Estimated eigenfunctions and eigenvalues corresponding to $\widehat{C}_{Ug}$ are denoted by $\widehat{\Phi}_k^{U_g} = [\widehat{\phi}_{ks}^{U_g}(t)], t \in \mathcal{T}, s = 1, .., P_{U_g}$ and $\widehat{\nu}_k^{U_g}$, where k is equal to the number of time points.

Step 4. Predict the random basis weights

Based on Karhunen-Loeve expansion, the random processes $\mathbf{U}_{gl}, g = 1, ..., G, l = 1, ..., L_{U_g}$ can be rewritten as $U_{gl} = \sum_{k=1}^{\infty} \xi_{lk}^{U_g} \Phi_k^{U_g}$. $\xi_{lk}^{U_g}$ are uncorrelated zero mean random basis weights (also denoted as FPC scores) with variance $\nu_k^{U_g}$. Thus, the random processes U_{gl} in model (2.9) will be replaced by the linear combination of FPC scores and eigenfunctions. Then, we can obtain the best linear unbiased predictors (BLUPs [4]) $\widehat{\xi}_{gl}$.

2.3 Clustering analysis

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). In simple words, the aim is to segregate groups with similar traits and assign them into clusters. Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms. In this section, I will introduce two clustering algorithms, K-means and self-organising map.

2.3.1 K-means

K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest distance. The algorithm can be divided into the following 3 steps.

Step 1. Find the local optimum for a specific k and initial means

For a specific k, the number of clusters and an initial set of k means $m_1^{(1)}, ..., m_k^{(1)}$, the algorithm iterates between the following two steps:

(Assignment step) Assign each point to the nearest mean. The distance is measured by Euclidean distance.

(Update step) After all data points are assigned, recalculate the mean of each cluster.

The algorithm converges until the assignments no longer change. But we should notice that the clustering result may not be the global optimaum for a specific k. Different sets of initial means may lead to different clustering results by above iterative steps.

Step 2. Find the global optimum for a specific k

The way to obtain the global for a specific k optimal one is to choose the one with minimum total variance within each cluster.

Step 3. Find the optimum k

Each time we add a new cluster, the total variance within each cluster is smaller than before. Thus, we can choose the k as the optimal k, after which the variance doesn't go down quickly. This is conducted by using "elbow plot" (reduction of variance VS number of clusters) and pick k by finding the "elbow" in the plot.

2.3.2 Self-organising map

The self-organising map (SOM) is an automatic data-analysis method. It is widely used to clustering problems. To illustrate what a SOM is, Let's introduce some notations and definations.

(1) a sequence of n-dimetional vectors $\{\boldsymbol{x}(t)\}$ represents input data items, where iteration $t = 1, ...T$ with T very large

(2) G is a lattice graph with grid nodes $\boldsymbol{v_i}, i = 1...n$

(3) $\{\boldsymbol{w_i}(t)\}$ is a sequence of n-dimensional weights $\{\boldsymbol{w_i}\}$, where i is the spatial index of the grid node with which $\{\boldsymbol{w_i}\}$ is associated.

(4) the neighborhood function $h(\boldsymbol{v_l}, \boldsymbol{v_k}, t) \in [0, 1]$ with $h(\boldsymbol{v_k}, \boldsymbol{v_k}, t) = 1$ and monotonically decreasing in terms of t and the distance $d(\boldsymbol{v_l}, \boldsymbol{v_k})$.

The relationship between (1)-(3) is that each input data is connected to grid nodes of a lattice graph G. Connections all have associated numbers called weights w.

The SOM algorithm is recursively determined by following equations

$$c = \underset{i}{\operatorname{argmin}} \|\boldsymbol{x}(t) - \boldsymbol{w_i}(t)\| \tag{2.11}$$

$$\boldsymbol{w_i}(t+1) = \boldsymbol{w_i}(t) + h(\boldsymbol{v_c}, \boldsymbol{v_i}, t)[\boldsymbol{x}(t) - \boldsymbol{w_i}(t)] \tag{2.12}$$

According to equation (2.11), the input data $\boldsymbol{x}(t)$ selects the best-matching weight (winner) associated with the node in G. Then, the weights of this node and those in its neighbor are modified according to equation (2.12) until the weights w_i doesn't change for each i [10]. Then we get stable weights $\boldsymbol{w} =: \{\boldsymbol{w}_i\}$. With index i of nodes $\boldsymbol{v_i}$ and the data $\{\boldsymbol{x}(t)\}, t = 1, ..., T$, we can form the cluster

$$\Delta_c(\boldsymbol{w}) = \{\boldsymbol{x}(t) : c = \underset{i}{\operatorname{argmin}} \|\boldsymbol{x}(t) - \boldsymbol{w_i}\|, \quad t = 1, ..., T\} \tag{2.13}$$

2.4 Spline

A spline is a special function defined piecewise by polynomials [21]. It provides a powerful tool for estimating nonparametric functions. In this section, I will introduce cubic spline interpolation used in data processing, and cubic B-spline which is chosen as basis in fitting functional linear mixed effect model.

2.4.1 Cubic spline interpolation

In practice, we often have a number of data points, obtained by sampling and experimentation, which represent a function that we don't know. In this case, it is often required to construct new data points within the range of known data points (knots). This process is called interpolation.

Consider we need to interpolate between all adjacent pairs of knots of $\{(x_i, y_i) : i = 0, ..., n\}$. Spline interpolation [21] uses low-degree polynomials in each of the intervals (x_{i-1}, y_{i-1}) and (x_i, y_i), and chooses the polynomial pieces $y = q_i(x), i = 1, ..., n$ such that they fit smoothly together. The resulting function is called a spline. The classical approach is to use polynomials of degree 3, called cubic splines, which can achieve the continuity of the first derivative and second derivative under the condition that the splines pass through all the knots. More specific, $y = q_i(x)$ should satisfy $q_i(x_i) = q_{i+1}(x_i)$, $q_i'(x_i) = q_{i+1}'(x_i)$ and $q_i''(x_i) = q_{i+1}''(x_i)$ for i=0,...,n-2.

2.4.2 B-spline

The term "B-spline" is short for basis spline. B-splines of order n are basis functions for spline functions of the same order defined over the same knots, meaning that all possible spline functions can be built from a linear combination of B-splines, and there is only one unique combination for each spline function [15]. B-splines are chosen as basis in my thesis because they have local scope. That is, the support of each individual B-spline is a closed interval so they only work locally.

Chapter 3

Apply FPCA-based FLMM to toxicity data

By observing the TCRC profiles in Figure 2, it is noticed that TCRCs from the same cluster may be quite different while some from different clusters have similar profiles [29]. Therefore, a proper way to capture correlation of profiles can not only achieve dimension reduction but increase clustering accuracy. Wavelet transform was demonstrated in Y.Zhang [29] to be a powerful tool for data compression and feature extraction. For the same purpose, a Linear Mixed Model based on Functional Principal Component Analysis (FPC-based FLMM) proposed by Jona Cederbaum in 2017 [2] is applied in our work. Functional Liner Mixed Model (FLMM) can account for both fixed effects and random effects, which capture different sources of variation by considering the deviation from population mean. While Functional Principal Component Analysis (FPCA) get data reduced based on as much as possible of the variation, by reconstructing the data using FPCs and scores from the model results, and comparing the reconstructed data with the original, we can demonstrate whether FPC-based FLMM is a good model to fit the data. The resulting scores can be regarded as a new lower-dimensional data set.

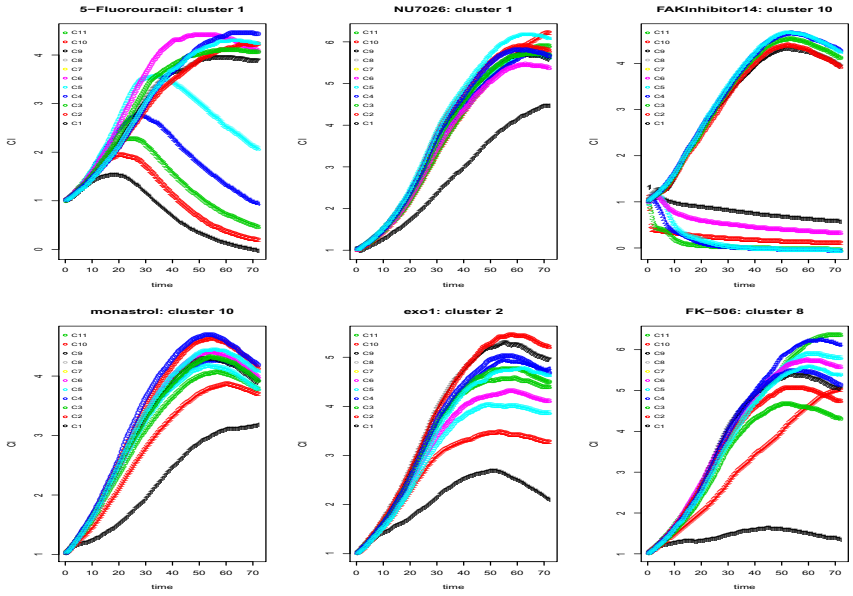


Figure 3.1: TCRCs from two different clusters

The same data was studied by Yongqing Yang in her report [26]. She clustered the 11 concentrations by K-means with the optimal number of clusters chosen by elbow method, and evaluated the clustering result by adjusted rand index. According to her report, the 11 concentrations are clustered into 3 levels with level1=(1,2,3,4), level2=(5,6,7,8), level3=(9,10,11). So, for the following demonstration, different concentrations from the same concentration level will be thought of as replications. For simplicity, "concentration level" is just termed as "concentration" in this thesis.

3.1 Model and model assumptions

There are multiple sources of variability in toxicity curves. First, there is variability between the different concentrations. Second, variability is also induced by different chemicals. Third, the repeated observations from the same chemical and same concentration induce variability. And fourth, there may be additional measurement error. Thus, we break down the variability induced by concentrations, chemicals and repetitions. FPC-based FLMM can allow us to decompose the variability in our data and to take advantage of the information on all sources of variability.

Since the repeated observations are nested within chemicals, and for each chemical we have measurements for each of the three concentrations, the FLMM here is a hierarchical model, which includes a fixed effect for concentration. The remaining hierarchy levels are accounted for by including random intercept. The model is in the following form

$$Y_{\tau co}(t) = \mu_\tau(t) + B_{\tau c}(t) + E_{\tau co}(t) + \epsilon_{\tau co}(t) \tag{3.1}$$

with $\tau = 1, ..., 3$ (concentrations), $c = 1, ..., n_c$ (chemicals), $o = \begin{cases} 1, ..., 4 & \tau = 1, 2 \\ 1, ..., 3 & \tau = 3 \end{cases}$ represents replicates. n_c is the number of chemicals used in the model. $Y_{\tau co}(t)$ represents the TCRC of concentration τ, chemical c, and replication o at time point t. $\mu_\tau(t)$ is the fixed effect for concentration. $B_{\tau c}(t)$ is a concentration-specific functional random intercept for chemicals. $E_{\tau co}(t)$ and $\epsilon_{\tau co}(t)$ are a smooth error term and white noise measurement error, respectively.

We assume that $B_{\tau c}(t)$ and $E_{\tau co}(t)$ are mutually uncorrelated random process with zero mean. Since the point-wise variations of concentrations differ from each other (Figure 2), we allow that the covariances of the chemical effects are different for each concentration, which is termed as "concentration-specific FRI". We assume the smooth error $E_{\tau co}(t)$ does not depend on concentration, thus is not concentration-specific.

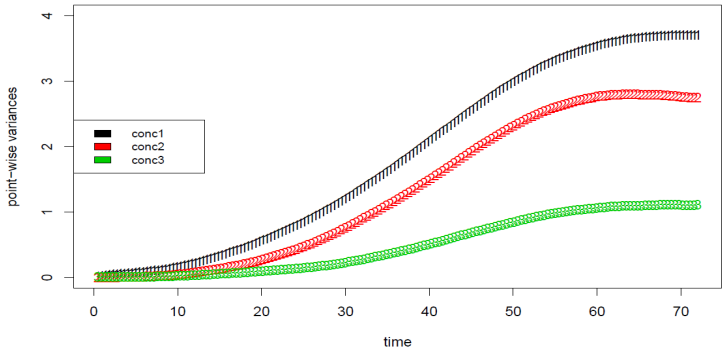


Figure 3.2: Point-wise variation curves per concentration

We denote the concentration-specific covariance of $B_{\tau c}(t)$ as $C^B_\tau(t, t')$, $(\tau = 1, ..., 3)$. The covariance of the smooth error $E_{\tau co}(t)$ is $C^E(t, t')$. The number of covariance to be estimated is 4 (denoted by G), which means we have 4 groups of random effects. The first three groups correspond to 1-3 concentration-specific chemical effects. The last group corresponds to the smooth error. For each group, the numbers of levels per group are $L_{U_1} = L_{U_2} = L_{U_3} = n_c$, $L_{U_4} = n_c * 11$. Then we specify one function random effect $P_{U_g} = 1$, g = 1,...,G = 4. So, there are $q = \sum_{g=1}^{G} L_{U_g} P_{U_g} = n_c * 14$ functional random effects in total.

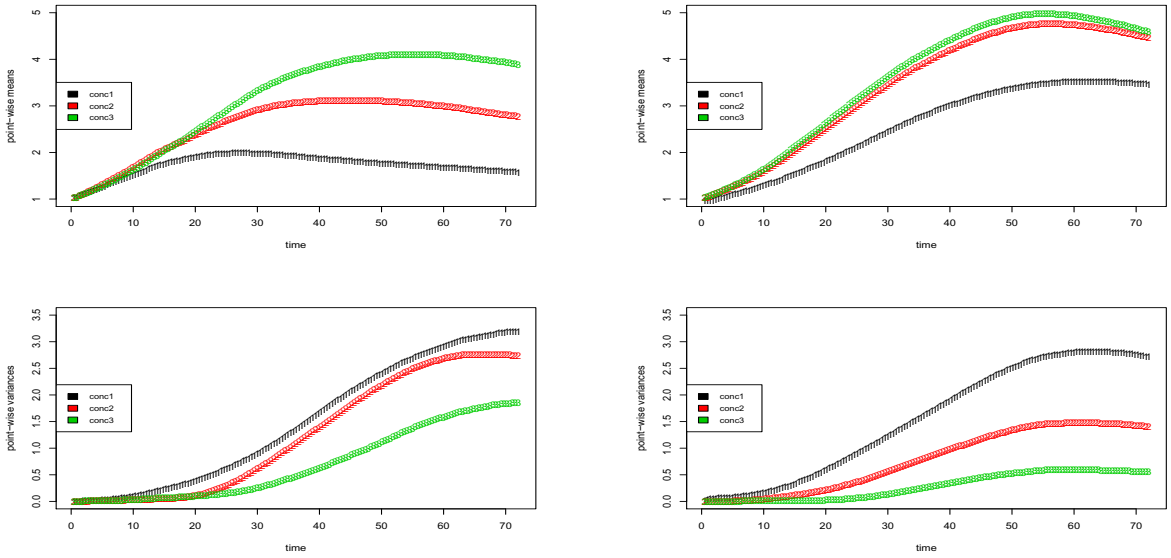
According to the demonstration above and model (6), the FREs part $B(t) + E(t)$ in model (8) can be represented in a matrix form ZU as following (assume applying the model to all the data:

65 chemicals)

$$B = \underbrace{\left[\begin{array}{c|c|c} \begin{matrix}1\\1\\1\\1\end{matrix} & & \\ \begin{matrix}\ddots\\ & 1\\ & 1\\ & 1\\ & 1\end{matrix} & & \\ & \begin{matrix}1\\1\\1\\1\end{matrix} & \\ & \begin{matrix}\ddots\\ & 1\\ & 1\\ & 1\\ & 1\end{matrix} & \\ & & \begin{matrix}1\\1\\1\\1\end{matrix} \\ & & \begin{matrix}\ddots\\ & 1\\ & 1\\ & 1\end{matrix} \end{array}\right]}_{\text{65 column}\quad\text{65 column}\quad\text{65 column}} \times \left[\begin{array}{c} U_{1,1} \\ \vdots \\ U_{1,65} \\ \hline U_{2,1} \\ \vdots \\ U_{2,65} \\ \hline U_{3,1} \\ \vdots \\ U_{3,65} \end{array}\right], \qquad E = \underbrace{\begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix}}_{65\times 11\text{ column}} \times \begin{bmatrix} U_{1,1,1} \\ \vdots \\ U_{1,65,4} \\ \vdots \\ U_{2,1,1} \\ \vdots \\ U_{2,65,4} \\ \vdots \\ U_{3,1,1} \\ \vdots \\ U_{3,65,3} \end{bmatrix}$$

3.2 Estimation results from R

Apply model on Cluster1 and Cluster10 Respectively

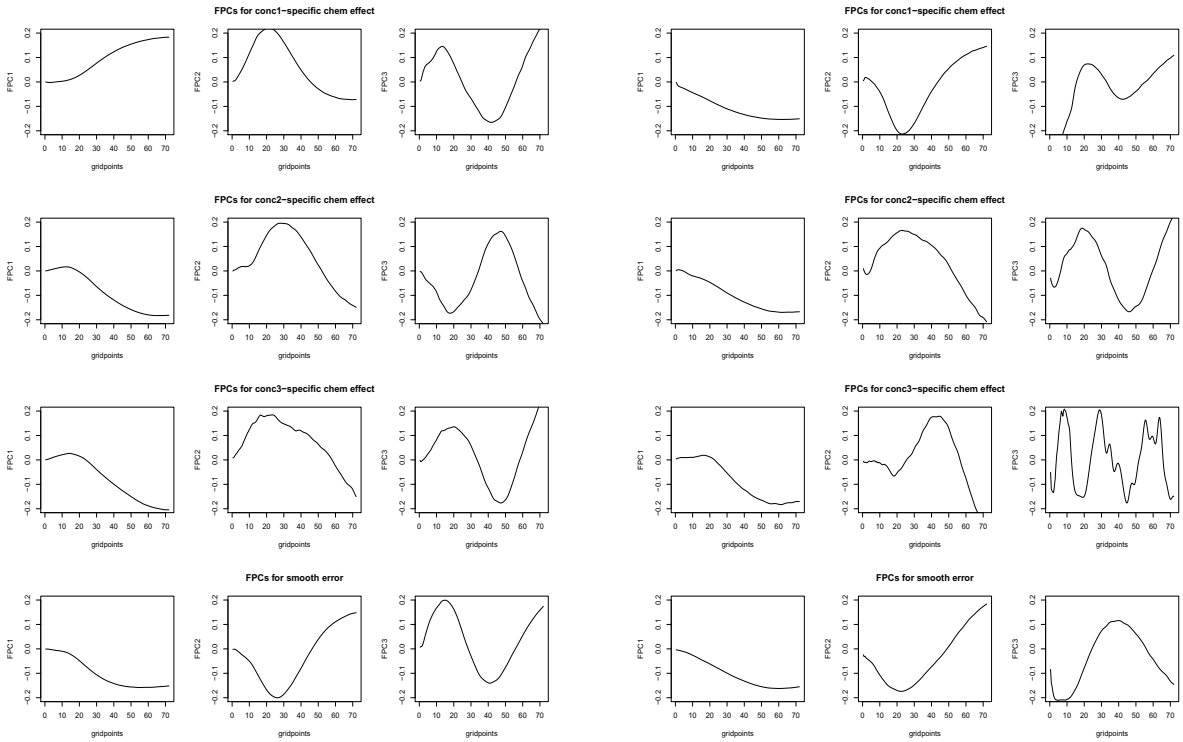


(a) Cluster 1

(b) Cluster 10

Figure 3.3: Point-wise mean curves for chemicals in cluster1 per concentration (upper left), Point-wise means for chemicals in cluster10 per concentration (upper right), Point-wise mean curves for chemicals in cluster10 per concentration (upper right), Point-wise variance curves for chemicals in cluster1 per concentration (bottom left), Point-wise variance curves for chemicals in cluster10 per concentration (bottom right)

As showed in Figure 3, point-wise mean curves become higher with the increase of concentration levels in both clusters. This is consistent with the fact that concentration 1 is the strongest which kills cells most efficiently. For both clusters, point-wise variance is largest of concentration 1, which implies in the case of high concentration, the increase in concentration will bring significant change in chemical efficacy, while it doesn't make much help to change the dose when the solution is not strong enough to kill cells.



(a) Cluster 1

(b) Cluster 10

Figure 3.4: FPCs for cluster1 chemical effects (left) and the cluster10 chemical effects (right): the first 3 FPCs of con1-specific effects (the first row), he first 3 FPCs of con2-specific effects (the second row), the first 3 FPCs of con3-specific effects (the third row), the first 3 FPCs of smooth error term (the fourth row)

The first and dominant FPC of each concentration is simple in structure, while the second and third FPCs have higher order. If we multiply the 3 FPCS in top right corner by -1, it is noticed that the first two of each concentration FPCs are similar between two clusters. The significant difference occurs in third FPC, especially for concentration 3, where the FPC in cluster10 has more cycles.

Apply model on all clusters

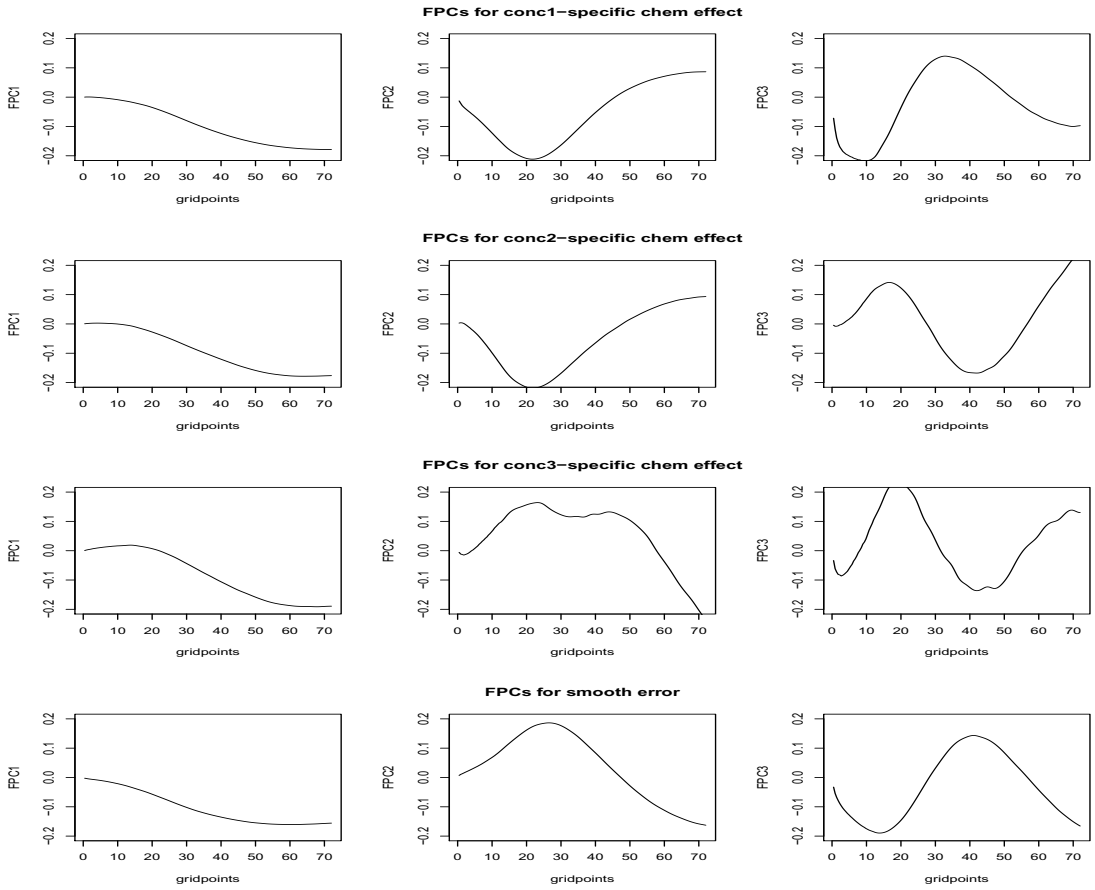


Figure 3.5: FPCs for all chemical effects: the first 3 FPCs of con1-specific effects (the first row), he first 3 FPCs of con2-specific effects (the second row), the first 3 FPCs of con3-specific effects (the third row), the first 3 FPCs of smooth error term (the fourth row)

The pattern of FPCs for all data are similar to those of cluster1 and cluster10 but the second and third FPCs are much smoother. The second FPC for concentration 3 has two peaks at around gridpoints 20 and 45, which seems like a combination of cluster1 and cluster10. The third FPC for concentration 3 is dominated by cluster1, which is consistent with the fact that sample size of cluster1 is largest.

The variability in each concentration can be decomposed into three sources as in Table 3.1. Within concentration 1, 69.2% variability is induced by chemical effects and 29.9% is by replication. The decomposition for concentration 2 is similar to concentration 1 with 64% variability explained by chemical and 29.9% explained by replication. In concentration 3, a large part of variability lies in replication while only 34.5% explained by chemical.

	variability source		
concentration	chemical	replication	measurement error
1	69.2%	29.9%	0.9%
2	64.0%	35.0%	1.0%
3	34.5%	63.7%	1.8%

Table 3.1: Variance decomposition

Then, we focus on chemical effects $B_{\tau c}$. The further decomposition of chemical effects on FPCs gives us interpretable measures of where in the TCRCs variability occurs between chemicals. For each concentration, the estimated first three principal components of $B_{\tau c}$ are depicted in Figure 3.6. For ease of interpretation, we show the effect of adding and of subtracting the estimated principal components multiplied by the square root of the respective eigenvalue to the concentration-specific mean. It can be considered as the information given by the corresponding direction (FPC).

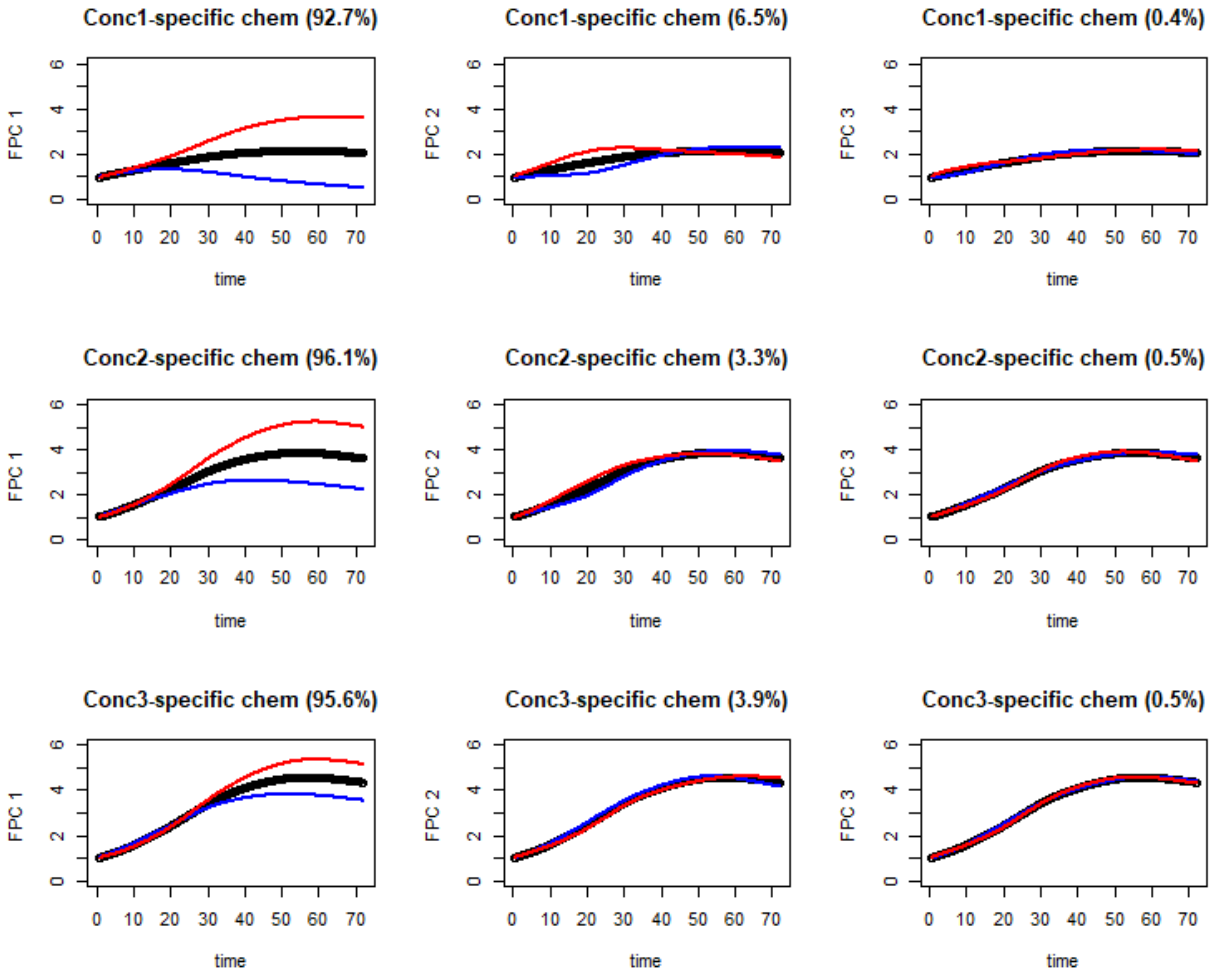


Figure 3.6: Concentration-specific means (black) plus (blue) and minus (red) chemical effects estimated by the first three components of concentration 1 (upper 3 figures), concentration 2 (middle 3 figures), concentration 3 (bottom 3 figures). The respective proportion of variability induced by chemical effects within concentration is given in brackets

The proportion in bracket represents the proportion of variability induced by corresponding concentration-specific chemical effects. The variability explained by the FPCs differs between concentrations. The first principal component of $B_{\tau c}$ explains most variability in concentration2-specific chemical effects (96.1%) followed by concentration 3 (95.6%). The first principal component of each concentration contain the information of overall variance in TCRCs. When the concentration is weakest (concentration 3), the chemicals take a long time to show difference. But for concentration 1 and 2, the difference of chemicals show up earlier, not because the red curve and the blue curve of the first principal component split away earlier at about 20 hours, but also the separation on the second principal component at the beginning. For concentration 1 and 2, we

see a vertical shift of the second principal on the mean curve after 40 hours. The third principal component in each concentration doesn't carry much information.

3.3 Reconstruct random effects

After having the conception of principal components and the information they carry, we will move on to visual presentation of random effects, chemical effects in Figure 3.7, 3.8 and replication effects in Figure 3.9, 3.10. In the following two figures, we just show several chemicals in concentration 1 from the two largest clusters as an example. The random effects depicted are obtained by the sum of first three principal components multiplied by the corresponding predicted scores.

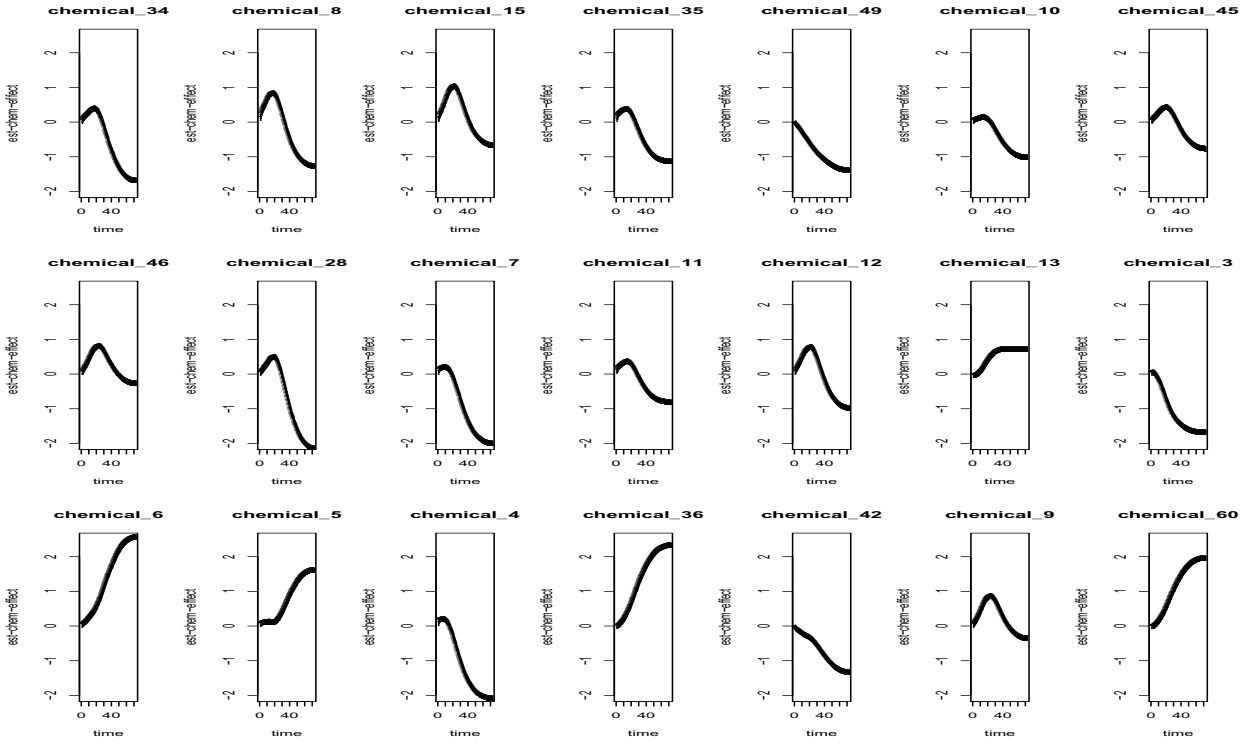


Figure 3.7: Conc1-specific chemical effects (cluster1)

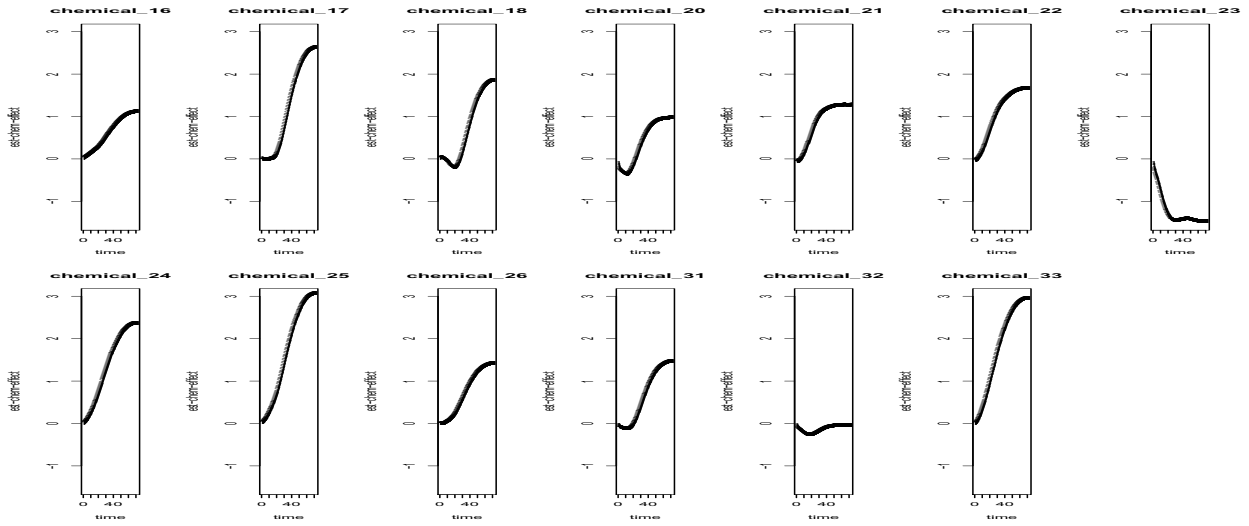


Figure 3.8: Conc1-specific chemical effects (cluster10)

In Figure 3.7, most conc1-specific chemical effects show the similar pattern-going up at the very beginning and going down afterwards with the slope becoming flat at the end. Chemical-13, chemical-6, chemical-5, chemical-36 and chemical 60 have different trend from others in cluster 1 but similar to common trend in Figure 3.8-an overall increase with some having a little decrease at the beginning. In Figure 3.8, only chemical-23 and chemical-32 are visually different from others in cluster 10, especially chemical-23, whose end point is much lower than starting point. All chemical effects can be thought as deviation from corresponding concentration-mean curve, which are the important foundation for the following chemical clustering.

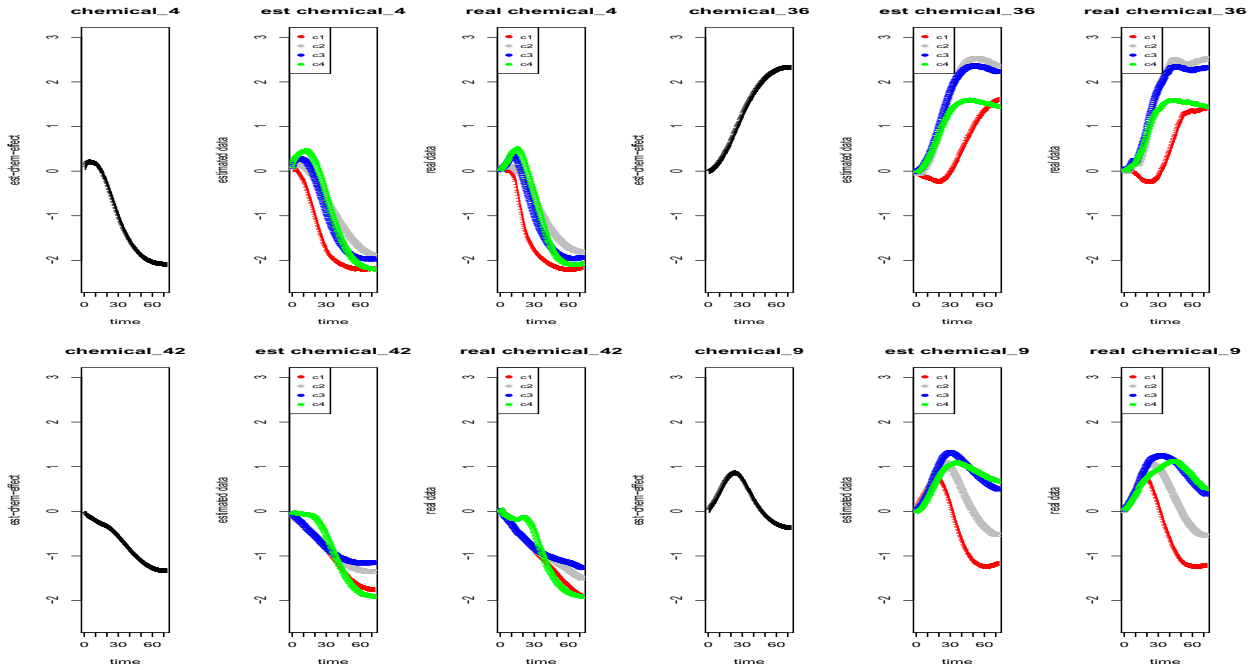


Figure 3.9: Real data VS estimated data of 4 chemicals (cluster1)

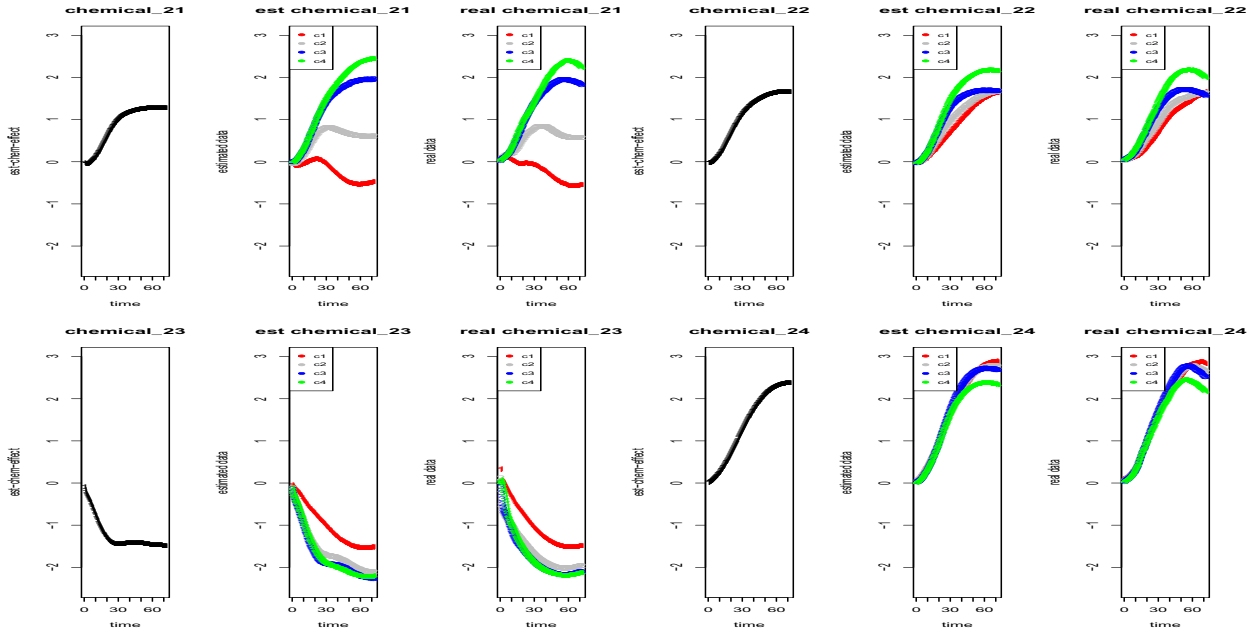


Figure 3.10: real data VS estimated data of 4 chemicals (cluster10)

Figure 3.9 and 3.10 display four chemicals in cluster 1 and cluster 10, respectively. For illustration, we use the three figures of chemical-4 (upper left) as an example. The first figure corresponds to "chemica_4" in Figure 3.7, which is predicted concentration1-specific chemical effect. The second

figure is the sum of random effects-replication effects added on chemical effect. The third figure is the raw TCRCs subtracting concentration1-mean curve. The legend "c1, c2, c3, c4" refer to 4 replications in concentration 1. By comparing "est" and "real" curves for each chemical, we notice that the truncated FPCs and corresponding scores chosen by model (3.1) and fit the data very well.

Also, comparison between predicted data and real data provide us a visible way to choose the number of eigen functions (see Figure 3.11 below). Using chemical-20 and chemical-26 in concentration 3 as examples, it is noticed that TCRCs predicted by the first two FPCs lose significant fluctuation while the first three FPCs capture more features of real TCRCs. That is why we choose 3 eigen functions as basis instead of 2.

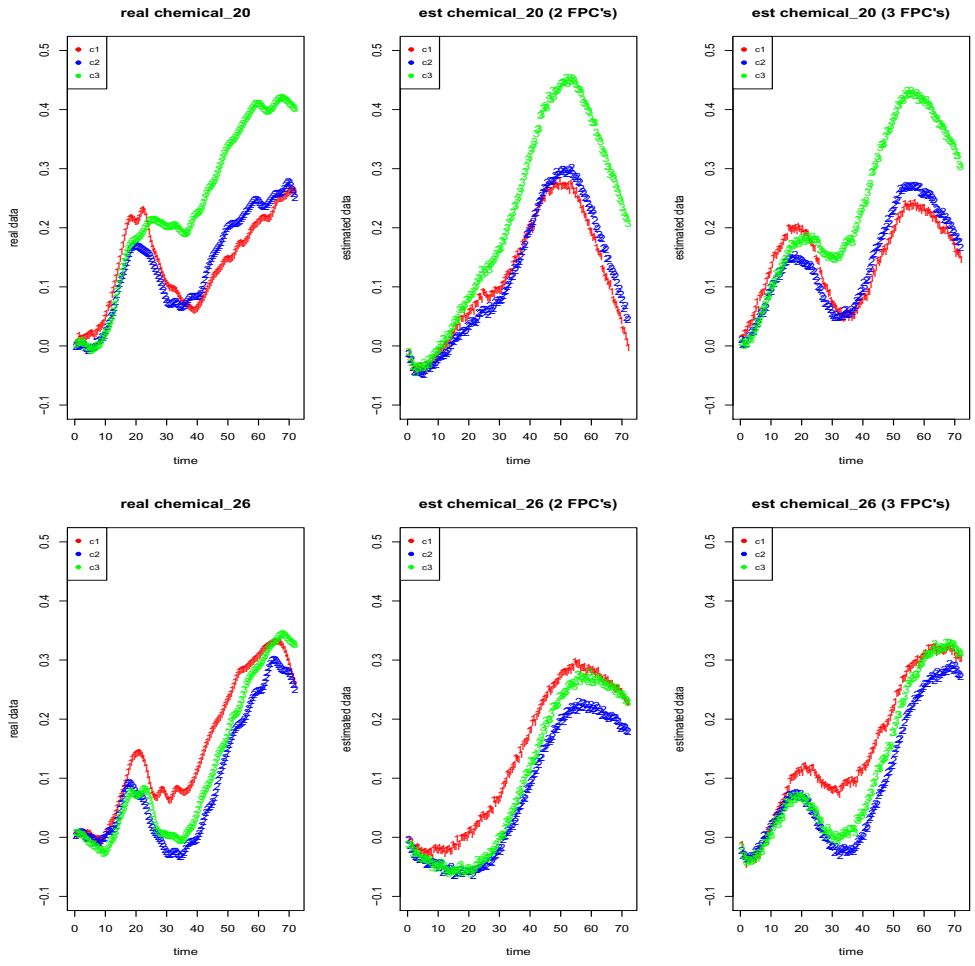


Figure 3.11: Chemical-20 in concentration 3 (the first row) and chemical-26 in concentration3 (the second row). For each row, the left figure is real data; the middle is estimated data by first 2 FPC's; the right is estimated data by first 3 FPC's

3.4 Binary clustering

We first consider the clustering of the two largest clusters, namely cluster 1 with target class DNA/RNA and cluster 10 with target class protein. There are 21 chemicals in cluster 1 and 13 chemicals in cluster 10.

3.4.1 Visual clustering by score plots

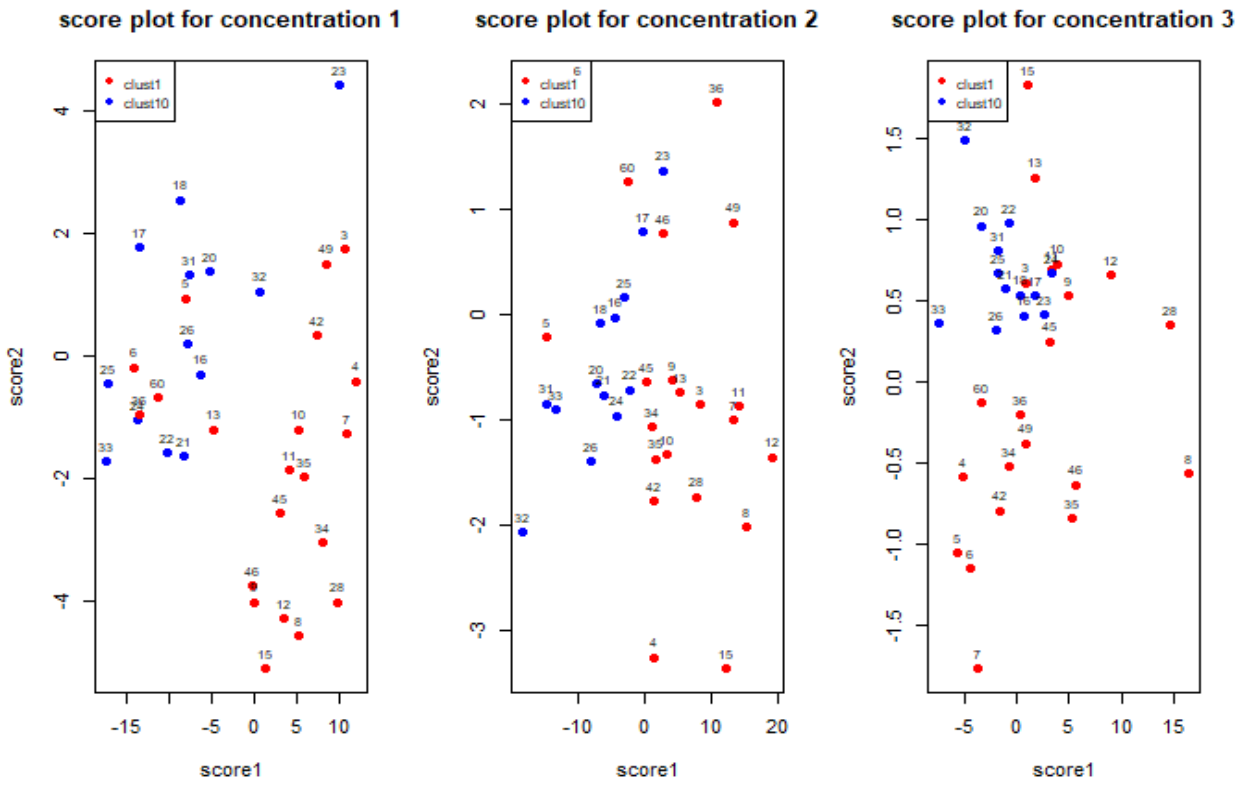


Figure 3.12: Score plots for 3 concentrations.

Score1 and score2 are eigen values corresponding to FPC1 and FPC2 of conc1-specific chemical effects (left). Score1 and score2 are eigen values corresponding to FPC1 and FPC2 of conc2-specific chemical effects (middle). Score1 and score2 are eigen values corresponding to FPC1 and FPC2 of conc3-specific chemical effects (right).Red points represent chemicals in cluster 1 and blue points are those in cluster 10.

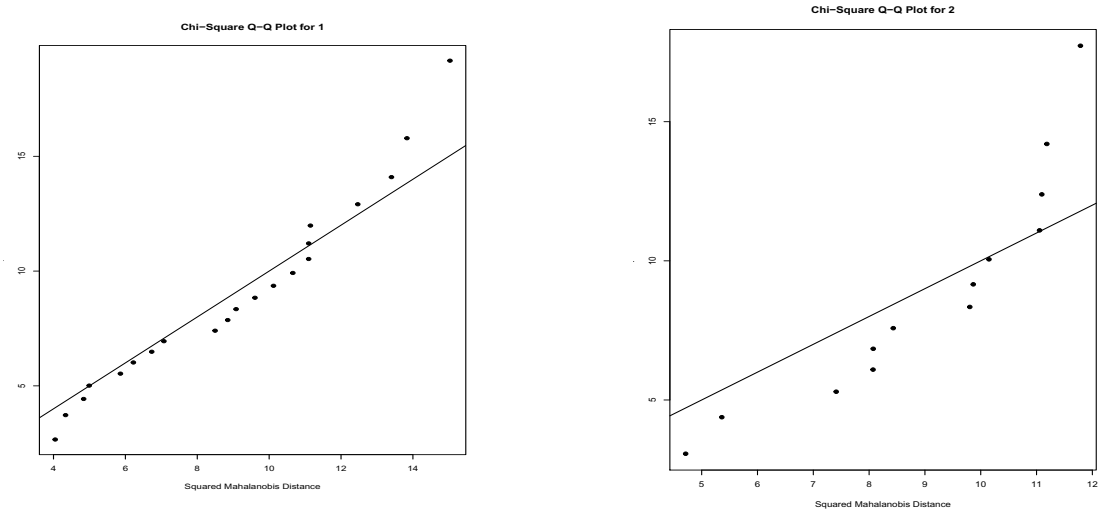
In the left figure, chemicals in two clusters are separated clearly except that chemical-5, chemical-6, chemical-13, chemical-36, chemical-60 are mixed in cluster 10 and chemical-23 is almost on the boundary of two clusters, which is consistent with what we noticed in Figure 3.7.

In the middle figure, FPC1 is dominant "direction" to separate the two clusters. Score1 of conc2-specific chemical effects in cluster 1 is positive except chemical-5, chemical-6, chemical-60. In cluster 10, score1 is negative with only chemical-23 positive. According to FPC1 for conc2-specific chemical effects in Figure 3.5 (the first figure in the second row), we can deduce that chemicals in cluster 1 have effect on killing cells with middle concentration level, while those in cluster 10 have the opposite effect on cells. In the right figure, chemicals in concentration 3 can not be split on FPC1. While on FPC2, chemicals in cluster 10 are all positive and they bunch together from 0.3 to 1.0 in terms of score2.

3.4.2 Test difference by Manova

As discussed above, we choose the first three FPCs for each concentration-specific chemical effects because they carry almost 100% variability induced from chemicals. For convenience, the scores of concentration 1 is termed as V1, V2, V3, the scores of concentration 2 as V4, V5, V6 and concentration 3 as V7, V8, V9. Thus, testing the difference between cluster 1 and cluster 10 is equivalent to testing if V1-V9 in cluster 1 are equal to V1-V9 in cluster 10 correspondingly, which is an One-Way Manova problem.

Before performing Manova, normality of data should be tested. As shown in Figure 3.13, normality assumption of data for two clusters seems plausible. Then, R function "MVN" in package "MVN" is applied to test multivariate normality of data. The results of normality tests also show that normality assumption is satisfied so that we can continue the following research.



(a) Cluster 1

(b) Cluster 10

Figure 3.13: Chi-square Q-Q plot

Unlike Anova in which there is only one test statistic (the F-ratio) to determine significance value, Manova provides us 4 different test statistics. As shown in table 1, the names of 4 statistics are "Wilks", "Pillai", "Hotelling-Lawley" and "Roy". The 4 test statistics are defined by Seber in 1984 [19] in terms of two matrices B and W and their degree of freedom, where B and W are sum of squares and cross-products matrices representing variance caused by treatment and error, respectively. "Wilks'Lambda" statistic proposed originally by Wilks corresponds to the equivalent form of the F-ratio in univariate case [8]. "Lawley-Hotelling Trace" and "Pillai's Trace" statistics are defined in terms of trace of BW^{-1} and $B(B+W)^{-1}$ respectively. "Roy's Largest Root" statistic is defined as the largest eigenvalue of BW^{-1}. Each test statistic can be constructed into an approximation, which is based on F-distribution and can be used to determine significance value. See Seber (1984) [19] for details.

test statistic	approx F	num DF	den DF	Pr(>F)
Wilks	0.25008	9	24	2.23e-05 ***
Pillai	2.9987	9	24	2.23e-05 ***
Hotelling-Lawley	0.74992	9	24	2.23e-05 ***
Roy	2.9987	9	24	2.23e-05 ***

Table 3.2: Manova

Since the number of treatments in our case is 2 (cluster1 and cluster10), degree of freedom of B is 1. According to Seber(1984) [19], all four test statistics will lead to identical results in such

case. So, the result of Manova (Table 3.2) based on 4 test statistics gives us the same result, which indicate there is significant difference between cluster 1 and cluster 10. More precisely, at least one pair of $V_i's, i = 1, ..., 9$ of cluster 1 and cluster 10 are different.

3.4.3 Clustering by k-means

As the input to Manova, we apply V1-V9 as input to K-means to do clustering. Firstly, we only choose the two largest clusters-cluster 1 and cluster 10.

chemical index	chemical	real cluster	clustering result
34	5-FU	1	1
8	gemcitabine HCl	1	1
15	gemicitabine	1	1
35	Etoposide	1	1
49	Doxorubicin	1	1
10	merbarone	1	1
45	Clofarabine	1	1
46	Hydroxyurea	1	1
28	SN38	1	1
7	Topotecanhydrochloride	1	1
11	irinotecan(CPT-11)	1	1
12	cytosine b-D-arabinofuranoside	1	1
13	ABT-888	1	1
3	Mitoxantronedihydrochloride	1	1
6	CRT0044876	1	2
5	NU7026	1	2
4	MitomycinC	1	1
36	Cordycepin	1	2
42	actinomycinD	1	1
9	cisplatin	1	1
60	OchratoxinA	1	2
16	monastrol	2	2
17	stritytl-cysteine	2	2
18	dimethylenenastrone	2	2
20	Y-27632	2	2
21	Ro32-3555	2	2
22	Batimastat	2	2
23	FAKInhibitor14	2	1
24	MLCKInhibPep18	2	2
25	PF573228	2	2
26	Blebbistatin	2	2
31	ML7 hydrochloride	2	2
32	HA1100 hydrochloride	2	2
33	PF431396	2	2

Table 3.3: Clustering result by k-means

In the column of "clustering result", the red ones are miss clustered chemical. Four chemicals of cluster 1 and one chemical of cluster 2 are miss clustered in total. Chemical-5, chemical-6, chemical-36 and chemical-60 are clustered outside from other chemicals of cluster 1. While chemical-23 of clustered 2 is miss clustered from others. The overall accuracy rate is 85.29%. Even though the accuracy rate is not low, the BSS/TSS given by K-means is only 57.8%, where BSS/TSS is basically a measure of the goodness of the clustering result K-means has found. SS obviously stands for Sum

of Squares, so BSS stands for "between-group deviance" and TSS stands for "total deviance". Ideally we want a clustering that has the properties of internal cohesion and external separation, i.e. the BSS/TSS ratio should approach 1. A small value of BSS/TSS and large values of BSS and TSS are probably induced from the sparsity of data in high-dimensional space. Thus, I try any possible subset of V1-V9 to look for better clustering results and good clustering results can give us information of the important directions where the difference between clusters lies in.

By using the subset of V1-V9 as input, a better clustering results is found. It is obtained by using V2 and V4 as input and the accurate rate is 88.2%. The value of BSS/TSS is a little bit higher (62.1%). By this way, we can also locate the significant difference between cluster 1 and cluster 10 lies in FPC2 of concentration1-specific chemical effects and FPC1 of concentration 2. In fact, only using V4 can achieve the same accurate rate to that from V1-V9. By observing score plot for concentration 2 in Figure 3.12, we can find that the points from cluster 1 and cluster 10 are well separated by score1=0, which is corresponding to V4. Combining the information in Figure 3.6, we can deduce that most chemicals in cluster 1 kill cells more efficiently than those in cluster 10. V2 added as input giving a better clustering rate indicates that some chemicals need concentration high enough to result in sharp change on cell population at around 20 hours, thus can be distiguishable.

3.4.4 Clustering by SOM

Using an exhaustive search through each of the SOM parameters, the effectiveness of the SOM's to segregate the data by MOA could be evaluated. Those results are presented in the following tables. The first four columns of the table denote which parameters were used for a given run of the SOM algorithm. Column five shows the overall accuracy rate.

neigh_func	topology	structure	grid	accuracy rate
bubble	hexagonal	Planar	6×5	79.41%
			4×3	88.24%
		Toroidal	6×5	88.24%
			4×3	88.24%
	rectangular	Planar	6×5	64.71%
			4×3	88.24%
		Toroidal	6×5	88.24%
			4×3	88.24%
gaussian	hexagonal	Planar	6×5	88.24%
			4×3	82.35%
		Toroidal	6×5	85.29%
			4×3	82.35%
	rectangular	Planar	6×5	67.65%
			4×3	82.35%
		Toroidal	6×5	82.35%
			4×3	82.35%

Table 3.4: Clustering results by SOM with different parameters

Table 3.5 shows a summary of the overall accuracy rate for each row shown above.

min	Q1	median	mean	Q3	max
64.71%	82.35%	83.82%	82.90%	88.24%	88.24%

Table 3.5: Accuracy rate summary

It is seen that changing the SOM parameters has some effect on the overall accuracy rate, with only two accuracy rates are below 70% and most of them are higher than 82.35%. The highest accuracy rate 88.24% is higher than that obtained from K-means. Then, we show one SOM clustering result in the 2-dimensional SOM plot in figure 3.14 with parameters being those in the second row in table 3.4.

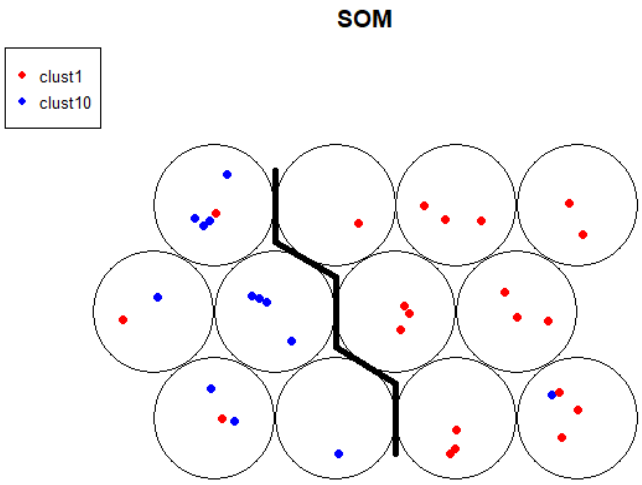


Figure 3.14: An example of clustering result by SOM

The separation is shown by the thick black line. We can see 3 chemicals in cluster 1 (red points) and 1 chemical (blue points) in cluster 10 are miss clustered.

Chapter 4

Apply FLMM with B-spline basis to toxicity data

In this chapter, I need to locate the time intervals where significant difference exists between clusters. Apart from cluster 1 and cluster 10, the number of chemicals in other clusters are very small. In order to avoid difficulties arising from such an imbalance among group size, I group all other clusters together as cluster 11. With the characteristic that different splines are dominant in different time intervals, cubic B spline is chosen as basis to fit linear mixed effect model.

According to figure 4.1, no matter which cluster or which concentration, the TCRCs go up at the beginning and go down later, some become stable at the end. The inflection points and the time becoming stable are different for different TCRCs. But most inflection points are before 20 hours or between 30 and 40 hours, while some after 50 hours. Overall, if we split the x-axis into 4 intervals evenly, it is enough to describe the main feature of the trend because there are not too many local wiggles along the TCRCs. Therefore, 4 cubic B spline basis functions are chosen to fit the TCRCs.

By comparing 3 clusters concentration by concentration in figure 4.1, we can see in concentration 1, cluster 1 and cluster 10 are relatively easy to distinguish. Most of TCRCs in cluster 1 starting going down at 20 hours, while TCRCs in cluster 10 never go down until the end, only becoming stable after 50 hours. The trend of TCRCs in cluster 11 is complicated, with some TCRCs are similar to those in cluster 1 and some are similar to cluster 10. In concentration 2, some TCRCs in cluster 1 have similar trend to those in cluster 10, only a few with significant inflection between 20 and 30 hours. Most of TCRCs in cluster 11 are similar to those in cluster 10. In concentration 3,

only 2 TCRCs in cluster 1 going down after reaching the peak, other TCRCs, no matter in which cluster, become stable at the end without any clear downward tendency in the whole process.

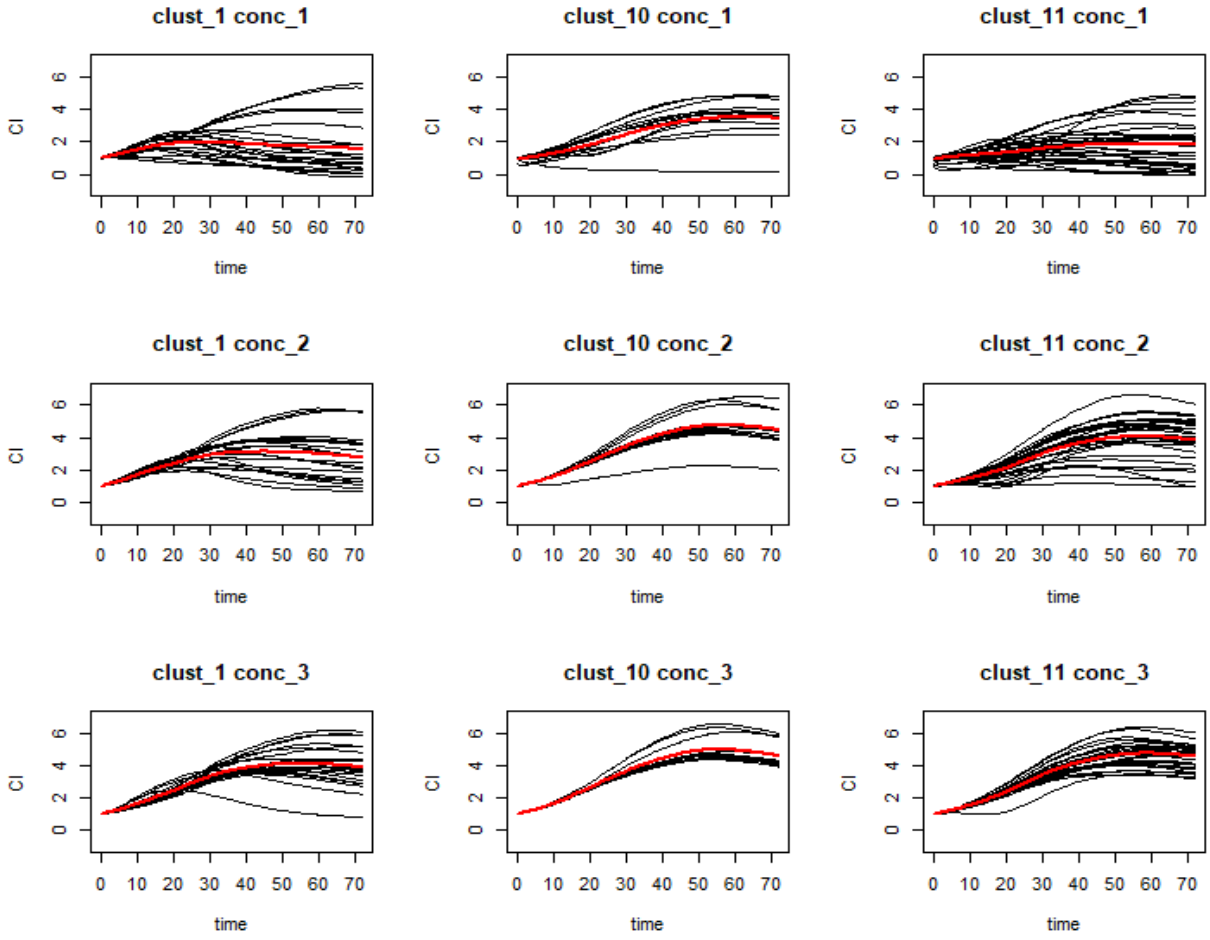


Figure 4.1: Each black curve corresponds to the average of a chemical TCRCs in a specific concentration level. The red curve is the average of black curves

4.1 Model and model assumptions

We consider the following model

$$Y_{\tau cok}(t) = \mu_{\tau k}(t) + B_{\tau c}(t) + E_{\tau co} + \epsilon(t) \tag{4.1}$$

with $\tau = 1, ..., 3$ (concentrations), $c = 1, ..., n_c$ (chemicals), $k = 1, 10$ represents clusters, n_c is the number of chemicals used in the model, $o = \begin{cases} 1, ..., 4 & \tau = 1, 2 \\ 1, ..., 3 & \tau = 3 \end{cases}$ represents replicates. $Y_{\tau cok}(t)$ represents the growth curve of concentration τ, chemical c's replication o and cluster k at time point t. $\mu_{\tau k}(t)$ is the fixed effect for cluster k concentration τ. $B_{\tau c}(t)$ is a functional random effect for concentration τ chemical c. $E_{\tau co}(t)$ is a functional random effect for concentration τ chemical c' replication o. $\epsilon(t)$ is white noise measurement error. The fixed effects and random effects are represented as the linear combination of spline basis functions as following

$$Y_{\tau cok}(t) = \sum_{i=1}^{4} \Big(\beta_{\tau ki}\phi_i(t)\Big) + \sum_{i=1}^{4} \Big(u_{\tau ci}\phi_i(t)\Big) + \sum_{i=1}^{4} \Big(w_{\tau coi}\phi_i(t)\Big) + \epsilon(t) \tag{4.2}$$

For each chemical c, we vectorize $Y_{\tau cok}(t)$, $\phi_i(t)$ and $\epsilon(t)$ in terms of t to get the following format

$$\mathbf{y}_{ck} = X\boldsymbol{\beta}_k + Z_1\boldsymbol{u}_c + Z_2\boldsymbol{w}_c + \boldsymbol{\epsilon} \tag{4.3}$$

where $\mathbf{y}_{ck} = \big(Y_{1ck}(t_1), ..., Y_{1ck}(t_n), ..., Y_{3ck}(t_1), ..., Y_{3ck}(t_n)\big)^T$. X, Z_1 and Z_2 are design matrices in terms of ϕ. $\boldsymbol{\beta}_k = \{\beta_{\tau ki} | \tau = 1, .., 3, k = 1, 10, i = 1, .., 4\}$ (unkown fixed). $\boldsymbol{u}_c = \{u_{\tau ci} | \tau = 1, .., 3, i = 1, .., 4\}$ (unknown random). $\boldsymbol{w}_c = \{w_{\tau coi} | \tau = 1, .., 3, \text{o} = \begin{cases} 1, ..., 4 & \tau = 1, 2 \\ 1, ..., 3 & \tau = 3 \end{cases}, i = 1, .., 4\}$ (unknown random). $\boldsymbol{u}$ and $\boldsymbol{w}$ are mutually independent.

4.2 Estimation results from R

Since cluster 11 is a combination of different small clusters, the chemicals in cluster 11 have different MOA, with some similar to those in cluster 1 and some similar to cluster 10. Multi-cluster clustering can not perform well in this case. Thus, we still apply binary clustering to cluster 1 and cluster 10, cluster 1 and cluster 11, cluster 10 and cluster 11 respectively.

Table 4.1 is the fixed effect summary of linear mixed model applied to data from cluster 1 and cluster 10. I use contrast to make cluster 1 as the baseline when I fit the model. In this table, the Intercept corresponds to the intercept of cluster 1 and Clust 10 corresponds to the difference between the intercept of cluster 1 and cluster 10. Meanwhile, Conc1:spline[, 1] represents the slope of cluster 1 concentration 1 on the first basis function , while Clust10:Conc1:spline[, 1] corresponds

fixed effects:	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	3.74510	0.28393	13.190	<2e-16 ***
Conc1:spline[, 1]	-2.79439	0.28525	-9.796	3.35e-16 ***
Conc2:spline[, 1]	-2.87917	0.28525	-10.094	< 2e-16 ***
Conc3:spline[, 1]	-2.88741	0.28575	-10.105	< 2e-16 ***
Conc1:spline[, 2]	-0.88395	0.34564	-2.557	0.011409 *
Conc2:spline[, 2]	-0.39924	0.34564	-1.155	0.249662
Conc3:spline[, 2]	-0.79934	0.34514	-2.316	0.021747 *
Conc1:spline[, 3]	-2.29368	0.49839	-4.602	8.09e-06 ***
Conc2:spline[, 3]	-0.24343	0.49839	-0.488	0.625858
Conc3:spline[, 3]	1.36286	0.50001	2.726	0.007072 **
Conc1:spline[, 4]	-2.11754	0.40142	-5.275	8.18e-07 ***
Conc2:spline[, 4]	-1.04028	0.40142	-2.592	0.011045 *
Conc1:spline[, 1]:Clust10	-0.04331	0.04424	-0.979	0.330145
Conc2:spline[, 1]:Clust10	0.04764	0.04424	1.077	0.284294
Conc3:spline[, 1]:Clust10	0.03985	0.04709	0.846	0.399053
Conc1:spline[, 2]:Clust10	-1.19533	0.31876	-3.750	0.000302 ***
Conc2:spline[, 2]:Clust10	-0.84400	0.31876	-2.648	0.009464 **
Conc3:spline[, 2]:Clust10	-0.27157	0.32055	-0.847	0.398932
Conc1:spline[, 3]:Clust10	2.85165	0.66240	4.305	4.03e-05 ***
Conc2:spline[, 3]:Clust10	2.63521	0.66240	3.978	0.000135 ***
Conc3:spline[, 3]:Clust10	1.39085	0.66327	2.097	0.038609 *
Conc1:spline[, 4]:Clust10	1.73893	0.45890	3.789	0.000264 ***
Conc2:spline[, 4]:Clust10	1.60637	0.45890	3.500	0.000706 ***
Conc3:spline[, 4]:Clust10	0.67481	0.45918	1.470	0.144938

Table 4.1: Summary of linear mixed effect model (cluster 1 VS cluster 10)

to the difference between the slope of cluster 1 concentration 1 on the first basis function and that of cluster 10. Reading the table this way, we can locate the significant differences between two clusters by focusing on Conc1:spline[, 2]:Clust10, Conc2:spline[, 2]:Clust10, Conc1:spline[, 3]:Clust10, Conc2:spline[, 3]:Clust10, Conc3:spline[, 3]:Clust10, Conc1:spline[, 4]:Clust10 and Conc2:spline[, 4]:Clust10.

We can get a preliminary conclusion that chemicals in cluster 1 and cluster 10 show more difference when the solution concentration level is at least 2. For both two concentration levels, the second, third and fourth basis functions are directions that correspond to cluster differences, which means, the significant difference happens after the first phase. According to the estimates of significant terms, we can see the coefficients on the second spline of cluster 10 are less than those of cluster 1, which means the cells grow faster in cluster 1 chemicals at the beginning. But on the last two splines, the coefficients of cluster 10 are much greater. This coincides with what we observed in figure 4.1 that chemicals in cluster 1 kill the cells more efficiently in the second half phase.

4.3 Reconstruct the TCRCs

I use the estimation $\widehat{\boldsymbol{\beta}}_k$, pridection $\widehat{\boldsymbol{u}}_c$ and model (4.2) to reconstruct the TCRCs. Figure 4.2 show three chemicals as examples. Black curves correspond to the average of replications of a specific concentration, while the colored curves are predicted TCRCs without using replication effects. It shows that linear mixed effect model with B spline basis perform well in capturing the overall feature of the TCRCs.

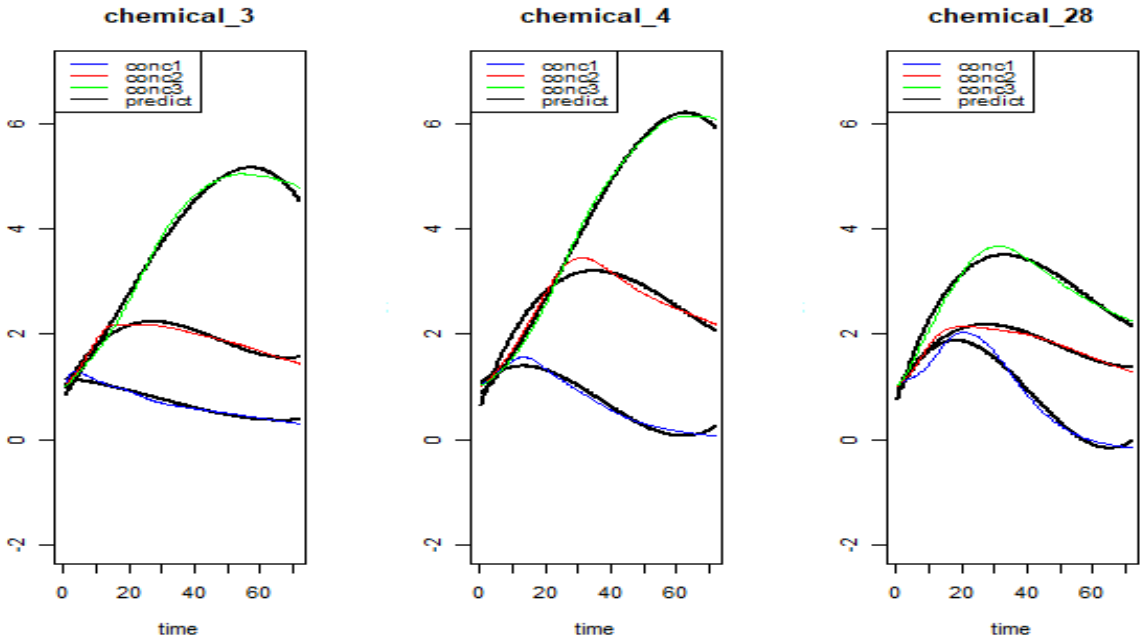


Figure 4.2: Raw TCRCs VS predict TCRCs

4.4 Binary clustering

Before doing clustering, we need to change the model (4.3) to the following model

$$\mathbf{y}_c = X\boldsymbol{\beta} + Z_1\boldsymbol{u}_c + Z_2\boldsymbol{w}_c + \boldsymbol{\epsilon} \tag{4.4}$$

The only difference between model (4.3) and (4.4) is that model (4.4) doesn't contain cluster indices k, thus $\boldsymbol{\beta} = \{\beta_{\tau i} | \tau = 1, .., 3, i = 1, .., 4\}$. Then, the coefficients of chemical effects are extracted as input to do clustering. For each chemical, the number of coefficients is 12, corresponding to the length of $\boldsymbol{u}_c = \{u_{\tau ci} | \tau = 1, .., 3, i = 1, .., 4\}$. By this way, the dimension of input is reduced.

4.4.1 Clustering by k-means

Cluster 1 VS cluster 10

By applying all possible combinations of components in $\boldsymbol{u}_c$ and comparing the clustering results, we find that the best clustering result is given by table 4.2, where 4 chemicals in cluster 1 and 1 chemical in cluster 10 are miss clustered and the overall accurate rate is 85.29%.

	1	2
clust1	17	4
clust10	1	12
Accurate Rate : 85.29%		

Table 4.2: Best clustering result (cluster 1 VS cluster 10)

Table 4.3 shows the three combinations as input that can obtain the best clustering result. The dimension in the table refers to the dimension of the input into K-means algorithm. The accurate rate is obtained by using the number of correctly-clustered chemicals divided by the total number of chemicals of cluster 1 and cluster 10. Even though the three combinations in table 4.2 give us the best clustering result, their values of BSS/TSS are different, where the first combination gives the highest value (74.9%). It means that when using the coefficients on the fourth spline of the TCRCs in concentration 1 and 2, K-means can give us the best result.

combination	dimension	accurate rate	BSS/TSS
Conc1:spline[, 4], Conc2:spline[, 4]	2	85.29% (29/34)	74.9%
Conc1:spline[, 2], Conc2:spline[, 4]	2	85.29% (29/34)	45.0%
Conc3:spline[, 3], Conc2:spline[, 4]	2	85.29% (29/34)	48.5%

Table 4.3: Different inputs in K-means and the corresponding results (cluster 1 VS cluster 10)

The following 2-dimensional plot of coefficients on fourth spline for concentration 1 and concentration 10 (figure 4.3) can demonstrate the separation clearly, where the points of cluster 1 are around the lower left corner but those of cluster 10 are in the upper right corner. Most coefficients of cluster 1 on fourth spline are less than 1, no matter for concentration 1 or concentration 2, whereas for cluster 10, they are greater than 1.

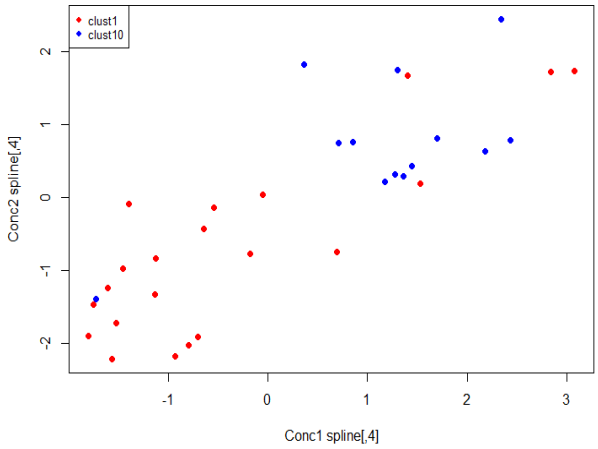


Figure 4.3: 2-dimensional plot of coefficients (cluster 1 VS cluster 10)

The explanation of figure 4.3 coincides with what we observed in figure 4.1 when comparing cluster 1 and cluster 10 concentration by concentration. For concentration 1, most TCRCs in cluster 1 end up going down to the starting level, while TCRCs in cluster 10 go up all the way and become stable above 2 at the end. For concentration 2, even though the TCRCs for the two clusters are much higher than those in concentration 1, the difference between the two clusters is still obvious in the last phase, with most of TCRCs in cluster 1 end up below 4 while in cluster 10 above 4. That is why the coefficients on the fourth spline for cluster 1 tend to be less than those for cluster 10. For concentration 3, the difference almost disappear since the concentration of the solution is too dilute to exert toxicity on cells. It is also noticed that even though the difference between the two clusters are obvious for most TCRCs , a few TCRCs in cluster 1 have the similar trend to cluster 10, and 1 TCRC in cluster 10 is far from others but is similar to cluster 1. That is why 4 chemicals in cluster 1 and 1 chemical in cluster 10 are miss clustered. The miss-clustered chemicals are the same to what we found in chapter 3, which are "CRT0044876", "NU7026", "Cordycepin" and "OchratoxinA" of cluster 1 and "FAKInhibitor14" in cluster 10.

Cluster 1 VS cluster 11

Similarly, the coefficients of all possible combinations of components in $\boldsymbol{u}_c$ of cluster 1 and cluster 11 are chosen as input to do clustering and the best clustering result is given by table 4.4, where 5 chemicals in cluster 1 and 2 chemicals in cluster 11 are miss clustered and the overall accurate

rate is 86.54%.

	1	2
clust1	16	5
clust11	2	29
Accurate Rate : 86.54%		

Table 4.4: Best clustering result (cluster 1 VS cluster 11)

Table 4.5 shows the four combinations as input that can obtain the best clustering result. According to the values of BSS/TSS, the two combinations don't have significant difference. We find that they have common terms Conc2:spline[, 2] and Conc3:spline[, 4], which means by using the coefficients on the second spline of the TCRCs in concentration 2 and fourth spline in concentration 3, K-means can give us the best result.

combination	dimension	accurate rate	BSS/TSS
Conc2:spline[, 2], Conc3:spline[, 4]	2	86.54%	38.7%
Conc2:spline[, 2], Conc3:spline[, 4], Conc3:spline[, 1]	3	86.54%	38.6%

Table 4.5: Different inputs in K-means and the corresponding results (cluster 1 VS cluster 11)

The following 2-dimensional plot of coefficients on second spline in concentration 2 and fourth spline in concentration 3 (figure 4.4) can demonstrate the separation clearly, where the points of cluster 1 are on the lower right side and those of cluster 11 are on the upper left.

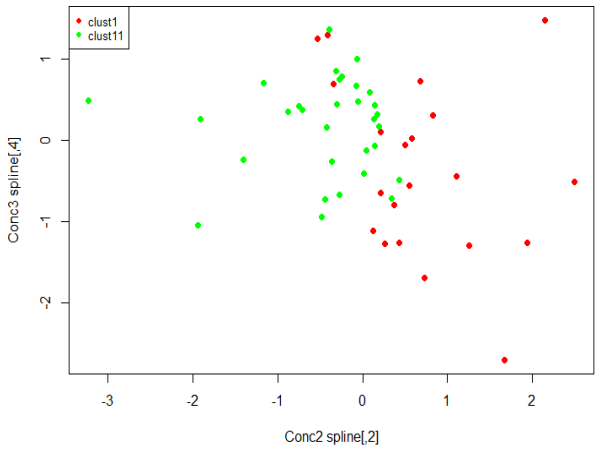


Figure 4.4: 2-dimensional plot of coefficients (cluster 1 VS cluster 11)

By comparing cluster 1 and cluster 11 concentration by concentration. Most TCRCs of the

two clusters have similar trend in concentration 1, which means when the concentration is strong, chemicals in the two clusters have similar effects on killing cells. In concentration 1, most TCRCs in the two clusters go up at the beginning and then go down until end up at round the starting level. For both clusters, some of the inflection points are a little bit earlier than 20 hours and some are a little bit later. In concentration 2, even though the overall trend of the two clusters are similar with increasing in the first half and then going a little down, the increasing patterns are different for the two clusters. The TCRCs in cluster 1 before 30 hours is concave down, while in cluster 11 is concave up. It means that the increasing rate of CI in cluster 1 is bigger than that in cluster 11, but it slows down later. Between 0 to 3 hours, most TCRCs in cluster 11 are below those in cluster 1. Thus, the coefficients of cluster 11 on second spline are much less than cluster 1. In concentration 3, the three clusters have similar trend because of the dilute solution of chemicals. However, at the very end, most TCRCs in cluster 1 are below 4 and more than half of the TCRCs in cluster 11 are above 4. Thus, adding the coefficients of the fourth spline in concentration 3 can increase the clustering rate.

Cluster 10 VS cluster 11

The best clustering result of cluster 10 and cluster 11 is given by table 4.6, where only 2 chemicals in cluster 10 and 6 chemicals in cluster 11 are miss clustered and the overall accurate rate is 81.82%.

	1	2
clust10	11	2
clust11	6	25
Accurate Rate : 81.82%		

Table 4.6: Best clustering result (cluster 10 VS cluster 11)

Table 4.7 shows the two combinations as input that can obtain the best clustering result. According to the values of BSS/TSS, the second combination is better. Actually, the dominant direction is Conc1:spline[, 4]. When only using the coefficients of Conc1:spline[, 4], the accuracy rate attains 79.5% and BSS/TSS is 72.5%.

combination	dimension	accurate rate	BSS/TSS
Conc1:spline[, 4], Conc2:spline[, 2]	2	81.82%	59.6%
Conc1:spline[, 4], Conc1:spline[, 3], Conc3:spline[, 3]	3	81.82%	62.6%

Table 4.7: Different inputs in K-means and the corresponding results (cluster 10 VS cluster 11)

The following 1-dimensional plot of coefficients on the fourth spline for concentration 1 (figure

4.5) can demonstrate the separation clearly, where the points of cluster 10 are on the right x-axis and those of cluster 11 are on the left. The boundary is around 0.5.

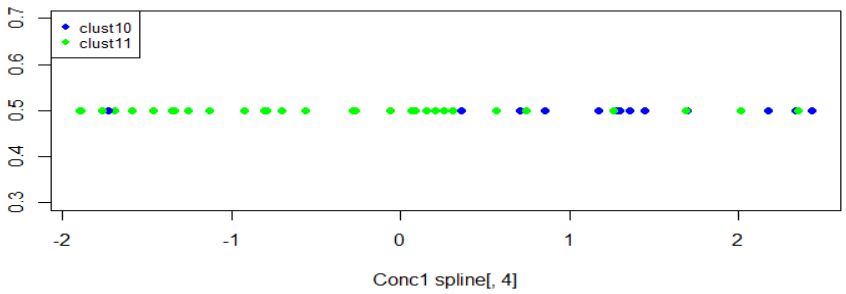


Figure 4.5: 1-dimensional plot of coefficients (cluster 10 VS cluster 11)

The explanation of figure 4.5 can be confirmed by comparing cluster 10 and cluster 11 concentration by concentration in figure 4.1. For concentration 2 and concentration 3, the two clusters have the similar patterns. For concentration 1, apart from several TCRCs in cluster 11 going up all the way, which are similar to those in cluster 10 thus miss clustered, others have totally different trend in the second half phase. TCRCs in cluster 10 go up a little after 35 hours and then become stable, but most TCRCs in cluster 11 go down after 35 hours, no matter the inflection happens at very early or upon the middle time. That is why the coefficients on the fourth spline for cluster 10 tend to be larger than most of those for cluster 11.

4.4.2 Clustering by SOM

When clustering TCRCs in cluster 1 and cluster 11, an exhaustive search is used through all possible combinations of SOM parameters and coefficients as input, and then the summary of the overall accuracy rate is obtained in table 4.8, where the "dimension of input" represents how many terms we choose as input to do clustering.

dimension of input	min	Q1	median	mean	Q3	max
1	50.00%	57.69%	61.54%	62.25%	67.31%	82.69%
2	50.00%	57.69%	65.38%	64.23%	69.23%	84.62%
3	50.00%	57.69%	65.38%	64.80%	71.15%	86.54%
4	50.00%	59.62%	65.38%	65.05%	71.15%	88.46%
5	50.00%	59.62%	65.38%	65.20%	69.23%	88.46%
6	50.00%	59.62%	65.38%	65.18%	69.23%	88.46%
7	50.00%	59.62%	65.38%	65.15%	69.23%	86.54%
8	50.00%	61.54%	65.38%	65.26%	69.23%	86.54%
9	50.00%	61.54%	65.38%	65.34%	69.23%	86.54%
10	50.00%	61.54%	65.38%	65.78%	69.23%	86.54%
11	50.00%	61.54%	65.38%	65.90%	71.15%	80.77%
12	55.77%	62.98%	65.38%	65.87%	69.71%	75.00 %

Table 4.8: Accuracy rate summary (cluster 1 VS cluster 11)

It is seen that the highest accuracy rate 88.46% is higher than that obtained from K-means. Among all combinations that can obtain the accuracy rate 88.46%, Conc2:spline[,2] is the common term. This coincides what we found in previous section that Conc2:spline[,2] is the most important direction where the significant difference between cluster 1 and cluster 11 lies in.

The similar exhaustive search is applied to clustering clusting cluster 1 and cluster 10, cluster 10 and cluster 11. The best accuracy rate is 91.18% for the former and 84.09% for the latter. Among all combinations that can obtain the best accuracy rate when clustering cluster 1 and cluster 10, Conc1:spline[, 2], Conc2:spline[, 3] and Conc2:spline[, 4] are the common terms, and when clustering cluster 10 and cluster 11, Conc1:spline[,4] is the common term. Those important directions are the same to what were found in previous section by using K-means.

The clustering results given by K-means and SOM and the corresponding inputs give us important information of distinguishing binary clusters. The interesting concentrations and time intervals are located in the process of model fitting and clustering. The MOA difference between chemicals in cluster 1 and cluster 10 lies in the last phase when using chemical solution with concentration 1 and 2. For cluster 1 and cluster 11, the best clustering can be obtained by applying the moderate concentration and the difference mainly exists in the first half phase. However, cluster 10 and cluster 11 are the most hard to distinguish, because cluster 11 is the combination of several small clusters and the MOA of some chemicals in it is more similar to that in cluster 10. Apart from those chemicals, the difference between cluster 10 and cluster 11 lies in the second half phase of when applying the strongest chemical solution.

4.5 Multi-cluster clustering

4.5.1 Clustering by k-means

The best clustering result for 3 clusters is given by table 4.9, where 5 chemicals in cluster 1, 1 chemical in cluster 10 and 8 chemicals in cluster 11 are miss clustered and the overall accurate rate is 78.46%.

	1	2	3
clust1	16	5	0
clust10	0	12	1
clust11	2	6	23
Accurate Rate : 78.46%			

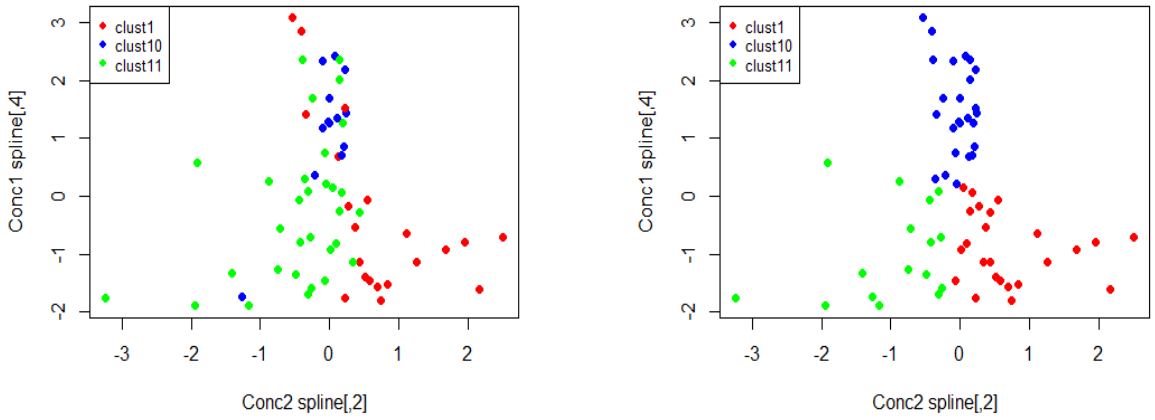
Table 4.9: Best clustering result by K-means (cluster 1 VS cluster 10 VS cluster 11)

Table 4.10 shows the combination as input that can obtain the best clustering result. Other combinations with more terms are based on the combination in Table 4.10 and when the dimension becomes higher, the value of BSS/TSS is smaller.

combination	dimension	accurate rate	BSS/TSS
Conc1:spline[, 4], Conc2:spline[, 2], Conc3:spline[, 4]	2	78.46%	58.5%

Table 4.10: Different input in Kmeans (cluster 1 VS cluster 10 VS cluster 11)

Since three dimensional plot is hard to observe the separation, we choose the dominant combination - Conc1:spline[,4] and Conc2:spline[,2], that can obtain 66.15% clustering rate. The following 2-dimensional plot of coefficients on the fourth spline for concentration 1 and second spline for concentration 2 (figure 4.7) can demonstrate the separation. According to figure 4.7 (a), most of the points of cluster 1 (red points) are around the lower right corner and the points of cluster 11 (green points) are around the lower left corner. Most points of cluster 10 (blue points) are in the middle in terms of x-axis and in the upper half of the y-axis with some points from the other two clusters mixed in this area. Conc2:spline[,2] is the most important direction that separates cluster 1 from the other two and the difference between cluster 10 and the other two mainly lies in Conc1:spline[,3]. Some points from cluster 1 and cluster 11 with much larger coefficients on Conc1:spline[,3] are hard to distiguish from cluster 10. By comparing the two plots, we can see many points in cluster 11 (green points) are miss clustered to cluster 1 and cluster10. Some miss clustered points are corrected by adding the third direction - Conc3:spline[,4] as input to do clustering.



(a) Real cluster

(b) Clustering result of K-means

Figure 4.6: 2-dimensional plot of coefficients (cluster 1 VS cluster 10 VS cluster 11)

The explanation of figure 4.7 coincides with the conclusions of binary clustering and what we observed in figure 4.1. In binary clustering setting, Conc2:spline[,2] is one of the most important directions that distinguishes cluster 1 and cluster 10 from cluster 11. Conc1:spline[,4] is one of the most important directions that distinguishes cluster 10 from cluster 1 and cluster 11. Con3:spline[,4] is one of the most important directions that distinguish cluster 1 and cluster 11, that is why some points miss clustered from cluster 11 to cluster 1 in figure 4.7 can be corrected by adding the coefficients of Con3:spline[,4]. In figure 4.1, for concentration 1, most TCRCs in cluster 10 go up all the way and become stable at the end while most TCRCs in cluster 1 and cluster 11 end up at about the starting level apart from a few TCRCs with trend similar to cluster 10. Thus, the points of cluster 10 are much higher in terms of Conc1:spline[,4] with some points from cluster 1 and cluster 11 mixed in. For concentration 2, a few TCRCs in cluster 1 and cluster 11 have obvious inflection points which cluster 10 doesn't have, but the overall trend of the three clusters are similar with increasing in the first half and then going a little down. However, in the first half phase, the increasing pattern of cluster 1 is different from the other two. Most TCRCs in cluster 1 before 30 hours are concave down, while in cluster 10 and cluster 11 are a little concave up. It means that the increasing rate of CI in cluster 1 is bigger than that in cluster 10 and cluster 11, but it slows down later. Thus, the coefficients on Conc2:spline[,2] of cluster 1 are larger than the other two. For concentration 3, most TCRCs in cluster 1 are below 4 at the end but more than half of the TCRCs in cluster 11 are above 4, even though when the concentration is strong enough. Thus, adding the

coefficients of Con3:spline[,4] can distinguish some TCRCs in cluster 1 and cluster 11 that are hard to separate when the concentration is strong enough.

4.5.2 Clustering by SOM

When clustering TCRCs of three clusters, an exhaustive search is used through all possible combinations of SOM parameters and coefficients as input, and then the summary of the overall accuracy rate is obtained in table 4.11.

dimension of input	min	Q1	median	mean	Q3	max
1	36.92%	46.15%	49.23%	49.27%	52.31%	64.62%
2	36.92%	47.69%	50.77%	51.47%	55.38%	72.31%
3	38.46%	49.23%	52.31%	52.55%	55.38%	76.92%
4	38.46%	49.23%	52.31%	53.00%	56.92%	75.38%
5	38.46%	49.23%	52.31%	53.35%	56.92%	75.38%
6	40.00%	49.23%	52.31%	53.55%	56.92%	76.92%
7	38.46%	49.23%	52.31%	53.71%	56.92%	73.85%
8	38.46%	49.23%	52.31%	53.71%	56.92%	75.38%
9	41.54%	49.23%	52.31%	53.56%	56.92%	72.31%
10	41.54%	47.69%	52.31%	53.21%	56.92%	73.85%
11	41.54%	47.31%	50.77%	52.55%	56.92%	70.77%
12	46.15%	47.31%	50.77%	52.12%	55.38%	64.62%

Table 4.11: Accuracy rate summary (cluster 1 VS cluster 10 VS cluster 11)

It is seen that the highest accuracy rate 76.92%. Among all combinations that can obtain the best accuracy rate, Conc1:spline[,4], Conc2:spline[,2] and Conc2:spline[,4] are the common terms. This almost coincides what we found in previous section by K-means. Then, we show an example of SOM clustering result in the 2-dimensional SOM plot in figure 3.14 with parameters that can obtain the best clustering result.

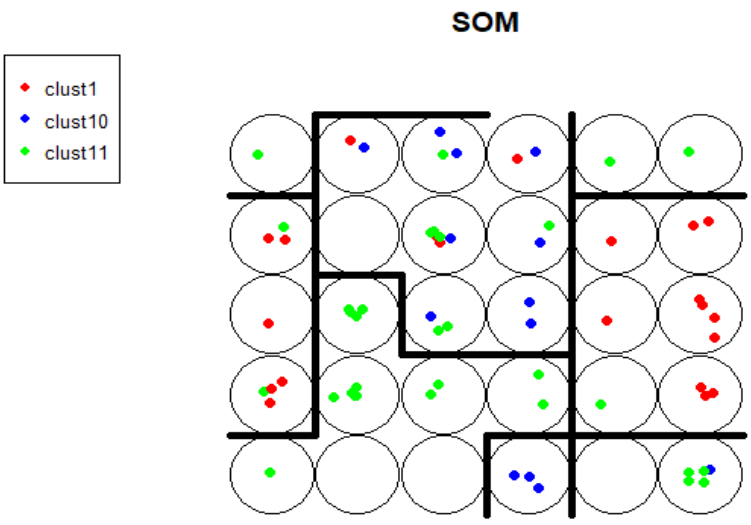


Figure 4.7: An example of clustering result by SOM

	1	2	3
clust1	17	4	0
clust10	0	12	1
clust11	3	7	21
Accurate Rate : 76.92%			

Table 4.12: Best clustering result by SOM (cluster 1 VS cluster 10 VS cluster 11)

The separation is shown by table 4.12 and the thick black line in figure 4.7. The big difficulty when doing multi-cluster clustering is that some chemicals in cluster 1 and cluster 11 have very similar MOA to those in cluster 10. Thus, 4 chemicals in cluster 1 and 7 chemicals in cluster 11 are miss clustered into cluster 10.

Chapter 5

Discussion and summary

A challenge of clustering functional data is that the observations are curves with distinct shapes, but not just points in Euclidean space. Thus, with limited sample size but high-dimensional data set, the results of clustering depend on what feasible model is chosen to fit the data. A good model can help clustering algorithms perform well by acting as a transformation of the curves that stretches the data in the direction corresponding to true cluster differences. The novelty of this paper is applying linear mixed effect model on the functional TCRCs before using traditional clustering algorithms, including a functional principal component based mixed model and a model with spline basis. Compared with linear model with only fixed effects, mixed model is able to effectively reduce the number of parameters need to be estimated by switching some factors from fixed effects to random. Therefore, linear mixed effect model is very suitable for our case as the effects of chemical within each concentration can be regarded as random deviation from the concentration mean. The two models both perform well in extracting direction information from different aspects to improve MOA clustering results.

The first linear mixed effect model takes advantage of functional principal component analysis, thus fully decomposes the variability induced from different sources by only using twelve FPCs as basis. Therefore, we can clearly know the amount of variation of chemicals and neglect the redundant information. A proper product of FPCs and corresponding eigenvalues can describe where in the TCRCs variability occurs between chemicals, including overall variability and local features. The process of utilizing the truncated FPCs to reconstruct the TCRCs gives a visible display of each chemical effect, which provides us a way to do preliminary clustering. The optimal binary clustering rate of cluster 1 and 10 from K-means and Self-organising mapping is 88.24%

by only using the scores on two FPCs. The dominant direction where the difference accurs is the first FPC of concentration2-specific chemical effects and the second FPC of concentration provides additional local information to distinguish some tricky chemicals. Carrying clustering on FPC basis is a useful way to determine the shape difference between clusters both in overall trend and critical local features.

The aim of fitting the other linear mixed effect model with spline basis is locating the time intervals where the difference occurs, since each spline function is dominant in a time interval. By specifying a cluster as reference level in the model, the directions with significant difference are located preliminarily. Then, by choosing the coefficients on any possible combinations of the significant directions to do clustering, the optimal clustering result and the corresponding directions can be determined by clustering accurate rate and BSS/TSS. Because group sizes for other clusters, apart from cluster 1 and 10, are much smaller, they are gathered as a third big group-group 11. For binary clustering, the accurate rate in the range of 79.55% to 100% is obtained. The important directions, on which clusters are distinguishable, are determined at the same time. Another impressive result is that multi-cluster clustering on spline basis performs well with accurate rate 73.85% and the important directions coincide with those in binary clustering. Overall, the MOA difference between chemicals in cluster 1 and cluster 10 lies in the last phase when using chemical solution in stronger concentration. For cluster 1 and cluster 11, the difference mainly exists in the first half phase of moderate concentration. The difference between cluster 10 and cluster 11 mainly lies in the second half phase of when applying the strongest chemical solution.

The MOA clustering results and the important directions found in the study have practical meaning in toxicity research. However, it is desirable to perform a reliable validation with more data available. In terms of future analysis, the information, especially the variance, from the linear mixed effect model can be applied to supervised methodology by constructing the distribution of known-cluster chemicals.

Bibliography

[1] BE Butterworth et al. "Chemically induced cell proliferation in carcinogenesis." In: *IARC scientific publications* 116 (1992), pp. 279–305.

[2] Jona Cederbaum. "Functional linear mixed models for complex correlation structures and general sampling grids". PhD thesis. lmu, 2017.

[3] L Fahrmeir et al. *Regression; Models, Methods and Applications. 2013*. 2003.

[4] David Harville. "Extension of the Gauss-Markov theorem to include the estimation of random effects". In: *The Annals of Statistics* (1976), pp. 384–395.

[5] Lajos Horváth and Piotr Kokoszka. *Inference for functional data with applications*. Vol. 200. Springer Science & Business Media, 2012.

[6] Keith A Houck and Robert J Kavlock. "Understanding mechanisms of toxicity: insights from drug discovery research". In: *Toxicology and applied pharmacology* 227.2 (2008), pp. 163–178.

[7] F Ibrahim et al. "Early determination of toxicant concentration in water supply using MHE". In: *Water research* 44.10 (2010), pp. 3252–3260.

[8] Richard Arnold Johnson, Dean W Wichern, et al. *Applied multivariate statistical analysis*. Vol. 5. 8. Prentice hall Upper Saddle River, NJ, 2002.

[9] Ian T Jolliffe and Jorge Cadima. "Principal component analysis: a review and recent developments". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065 (2016), p. 20150202.

[10] Teuvo Kohonen. "Essentials of the self-organizing map". In: *Neural networks* 37 (2013), pp. 52–65.

[11] Ping Ma and Wenxuan Zhong. "Penalized clustering of large-scale functional data with multiple covariates". In: *Journal of the American Statistical Association* 103.482 (2008), pp. 625–636.

[12] TH Pan et al. "Recognition of chemical compounds in contaminated water using time-dependent multiple dose cellular responses". In: *Analytica chimica acta* 724 (2012), pp. 30–39.

[13] Tianhong Pan et al. "Cytotoxicity assessment based on the AUC50 using multi-concentration time-dependent cellular response curves". In: *Analytica chimica acta* 764 (2013), pp. 44–52.

[14] Tianhong Pan et al. "In vitro cytotoxicity assessment based on KC50 with real-time cell analyzer (RTCA) assay". In: *Computational biology and chemistry* 47 (2013), pp. 113–120.

[15] Hartmut Prautzsch, Wolfgang Boehm, and Marco Paluszny. *Bézier and B-spline techniques*. Springer Science & Business Media, 2013.

[16] Mark Raffray et al. "Apoptosis and necrosis in toxicology: a continuum or distinct modes of cell death?" In: *Pharmacology & therapeutics* 75.3 (1997), pp. 153–177.

[17] JO Ramsay and BW Silverman. "Principal components analysis for functional data". In: *Functional data analysis* (2005), pp. 147–172.

[18] JO Ramsay and BW Silverman. "Springer series in statistics". In: *Functional data analysis* (2005), pp. 10–18.

[19] GAF Seber. "Multivariate observations john wiley & sons". In: *New York* (1984).

[20] H Slanina et al. "Real-time impedance analysis of host cell response to meningococcal infection". In: *Journal of microbiological methods* 84.1 (2011), pp. 101–108.

[21] Peter Turner. *Numerical Analysis*. London: Macmillan Education UKImprint Palgrave, 1994. ISBN: 9781349131082.

[22] Yuedong Wang. "Mixed effects smoothing spline analysis of variance". In: *Journal of the royal statistical society: Series b (statistical methodology)* 60.1 (1998), pp. 159–174.

[23] Zhankun Xi et al. "Mode of action classification of chemicals using multi-concentration time-dependent cellular response profiles". In: *Computational biology and chemistry* 49 (2014), pp. 23–35.

[24] Zhankun Xi et al. "Mode of action classification of chemicals using multi-concentration time-dependent cellular response profiles". In: *Computational biology and chemistry* 49 (2014), pp. 23–35.

[25] James Z Xing et al. "Microelectronic cell sensor assay for detection of cytotoxicity and prediction of acute toxicity". In: *Toxicology in vitro* 20.6 (2006), pp. 995–1004.

[26] Yongqing Yang. "Functional data clustering: A comparison of self-organizing maps and K-means in toxicity assessment".

[27] Ming Zhang et al. "Measuring cytotoxicity: a new perspective on LC50". In: *Anticancer research* 27.1A (2007), pp. 35–38.

[28] Ming Zhang et al. "Predicting tumor cell repopulation after response: mathematical modeling of cancer cell growth". In: *Anticancer research* 26.4B (2006), pp. 2933–2936.

[29] Yile Zhang et al. "Machine learning algorithms for mode-of-action classification in toxicity assessment". In: *BioData mining* 9.1 (2016), p. 19.