# MULTIVARIATE TIME SERIES CLUSTERING USING KERNEL VARIANT MULTI-WAY PRINCIPAL COMPONENT ANALYSIS

by

HWANSEOK CHOI

A DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Applied Statistics
in the Graduate School of
The University of Alabama

TUSCALOOSA, ALABAMA

2010

# ABSTRACT

Clustering multivariate time series data has been a challenging task for researchers since data has multiple dimensions to consider such as auto-correlations and cross-correlations whereas multivariate time series data has been prevailing in diverse areas for decades. However, for a short-period time series data, conventional time series modeling may not satisfy the model validity. Multi-way Principal Component Analysis can be used for this case, but the normality assumption can restrict to handle nonlinear data such as multivariate time series with high order interactions. Kernel variant MPCA will be proposed for an alternative solution for this case.

To test if KMPCA can cluster trivariate time series data into two groups, two simulation studies were conducted. The first study has the same mean structure groups with error structures which are combinations of three different auto-correlation levels and three different cross-correlation levels. Two different mean structure groups with nine error structures were generated for the second study. To check the proposed method work well on a real-world data, Obesity-depression relationship study was done for a real-world data.

The simulation studies showed that KMPCA cluster two different mean structure groups over 90% success rates when an appropriate kernel function with proper parameter was applied. Similar error structure will obstruct the clustering performance: strong cross-correlation, weak auto-correlation, and larger number of temporal points. Considering racial effect, obesity and obesity related variables, especially addictive material uses for 15 years can expect depressed cohorts at year 20 up to 76% for Caucasian group and 95% for African-American group.

ACKNOWLEDGMENTS

It is a pleasure to thank those who made me through this journey. Most of all, I would like to thank my advisor, Dr. J. Michael Hardin. This dissertation research would not have been done without Dr. Hardin. He always encouraged me whenever I had hard times with different, various obstacles I met. I truly appreciate Dr. Hardin for his support and friendship.

I also gratefully appreciate all committee members: Dr. Michael Conerly always listened to and helped me to see different ways when I was wandering in the corner. Dr. Brian Gray's insightful comments and constructive criticisms at different stages of my research were thought-provoking. I am grateful to him for holding me to a high research standard and enforcing strict validations for each research result, and thus teaching me how to do research. I am so appreciate Dr. Samuel Addy. I could not come this far without his support and encouragement. He has been not only my friend but also my mentor. Dr. Lee is one of the best teachers that I have had in my life. I am indebted to him for his continuous encouragement and guidance.

I thank Dr. Sharina Person who helped me to find real data and made it real for the case study. I also thank Dr. Young-il Kim, Dr. Polly P. Kratt, and Lucia Juarez for being supportive friends as well as understanding supervisors.

Most importantly, none of this would have been possible without the love and patience of my family. I would like to express my heart-felt gratitude to my family. I thank my parents, Hyoung Taik Choi, Ok-Im Lee, and my parents-in-law, Chang-sam Jin, Jaebun Jung, for being supportive parents.

Last, but not least, I would like to thank my wife, Soyoung for her understanding and love. Her support and encouragement was in the end what made this dissertation possible. And, my lovely children, Erin, Ethan, and Ean, without you all, I could not even take a step everyday. I love you so much.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

# CHAPTER 1

# INTRODUCTION

## 1.1 Background

Suppose one has to decide on a treatment plan for a patient from another patient or from a group of patients with the patient's physical records measured in three-hour intervals for several days. Or if one has to make a decision on a potential credit customer whose credit records, such as transaction history cover too short a time period, the decision would not be an easy task. In an industrial plant, an engineer could obtain subsequent data that contain information about the behavior of a certain machine process. It would be beneficial if these data could be categorized into groups of conditions, for example, severe symptoms or not; good credit or default; and normal or contaminated process, etc. Then, these characteristics could be used to support decisions in diagnosis, default detection, or error detection.

There has been a lot of research on clustering time series data in various areas, including science, engineering, business, finance, economics, health care, and government. However, it is a challenging task for researchers since these temporal data has auto-correlation, one of the time series data characteristics. In addition to that, if one has to deal with more than two variables which are more or less correlated with each other, then the job would be even more complicated.

Conventionally, modeling a time series is the most popular method in time series analysis. Once the time series data is modeled, clustering with the coefficients of the model is performed well enough. However, if there are not enough data points as in the cases mentioned above, it is not guaranteed that a valid model will be obtained from a given data at hand. Also, without any prior information, building a model does have limitations. Thus, exploring the data by clustering is a more appropriate method than modeling.

## 1.2 Clustering

Clustering is grouping objects into different groups or partitioning a data set into subsets, where a subset has strongly similar features among its members but dissimilar features from other clusters. Thus, it is one of the rudimentary, exploratory procedures to understand the complex nature of relationships between data, experiment subjects, and variables especially in the multivariate case. Without any prior information, this method can provide an informal means of assessing dimensionality, identifying outliers, and suggesting interesting hypotheses concerning relationships (Johnson and Wichern 2002, chap. 12).

Conventionally, clustering methods have been developed and performed for static data where all features do not change with time; they can largely be categorized into five methods: partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods (Liao 2005). These methods are for different contexts and various data sets; thus, a specific method works better to group objects with certain data characteristics than any other method does.

One of the important components is the criterion to measure the similarity between objects. It could be a distance measure such as Euclidean distance, root mean square distance, and Mahalanobis distance. Instead of raw data, a feature or an index of data could be compared.

Vectors or matrices could be compared also. Therefore, the method of measure similarity could vary depending on the characteristics of the objects to be compared as in the clustering method.

Another key component of clustering is the evaluation method after clustering. We can differentiate the case that the true groups are already known from the case that they are not known yet. When it is not known, it is harder to tell how well a certain clustering method works. Information criteria such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) could be used for this.

### 1.3 Time series

A time series is a sequence of real numbers, each number representing a value of an attribute of interest, $x_t$, each one being recorded at a specified time $t$. Thus, an observation $x_t$ depends on the adjacent time points, which is called auto-correlation. It restricts the applicability of the many conventional statistical methods that depend on the assumption that the observations are independent and identically distributed.

Since the early 1900s when one of the earliest time series, the monthly sunspot number series studied by Schuster in 1906, was recorded, from economics to social sciences, to physical and environmental sciences, to industrial batch processes, and to more recently speech recognition, time series data and analysis have been pervasive in the research world in disciplines ranging (Shumway 1988). In many cases, subsequent data for an object provides a better explanation of the current status of the object or of the status expected in the close future.

For example, when a potential cardiovascular patient is admitted to a hospital, he or she could be diagnosed using not only current data but also recent physical data. When it comes to credit scoring, a potential credit customer could be better explained by his or her credit history than by just status quo.

## 1.4 Time series analysis

Analysis of time series data could be categorized into identifying patterns which describes characteristics of the time series such as trend, seasonality, cyclical patterns, and irregularities as well as forecasting, which try to expect the next move of a certain time series in the future. A trend is defined by upward, downward, or deterministic movement in a time series. In this sense, deterministic implies that the movement can be approximated by a mathematical function which could be linear, nonlinear, or curvilinear. Trends can sometimes be modeled as a mathematical function or differencing. Seasonal factors can be explained by the fact that a time series repeats at one or more seasonal periods. This could be handled by seasonal differencing. Cyclical behavior is a repetitive pattern that depends not on seasonal factors but on economic or other factors. Irregular components of a time series result in a series that cannot be accounted for by known factors such as trends, seasonal or cyclical behaviors, a composition of variables, or any pattern. So, it could be called random error, white noise, or innovation. In sum, a time series can be decomposed into a model structure and an error part. If one can distinguish a specific pattern from the data, then one could describe, explain, predict, and control the cases.

Before looking at the methods of modeling a time series, it is necessary to know the concept of stationarity. A time series may have regularity over time in its behavior. That is, there exists a certain pattern or a structure which is obviously consistent for the time being. Once one figures out this pattern, one can formalize this with a function. To this end, it is necessary to assume: (1) the mean does not change over time, (2) the variance of the series does exist, and (3) the autocorrelation is the same at the pairs of points separated by a certain interval. It is called a weak stationary time series when these three assumptions are satisfied. This concept is critical for parameter estimation and forecasting in time series analysis.

Generally, it is appropriate to see that there are two separate but not necessarily mutually exclusive approaches to time series analysis, commonly identified as the time domain approach and the frequency domain approach (Shumway 1988).

The frequency domain approach assumes that the time series is best regarded as a sum or linear superposition of periodic sine and cosine waves of different periods or frequencies. This method deals with the autocorrelation generated in stationary time series by transforming a series to the frequency domain where adjacent observations are nearly independent. Through the finite version of Fourier transformation, transformed data are almost uncorrelated, approximately normally distributed, and have variances equal to the power spectrum. However, for the valid spectral approximation, it is inevitable to obtain quite a large data set. Compared to the time domain approach, it is more complicated to understand the analysis.

The time domain approach is implied by the assumption that correlation in adjacent time series values is best explained in terms of linear difference equations with constant coefficients. This method considers estimation of the coefficients of linearly related former values and error terms and determination of how many linear terms are needed. Box and Jenkins (1976) formalized the theory and methodology for time series forecasting. According to various sources of correlation, the model could be an autoregressive (AR) model and an autoregressive moving average (ARMA) model. If one would like to model nonstationary time series using differencing operators, an autoregressive integrated moving average (ARIMA) model is appropriate (Enders 1995). This time domain approach is much easier and more understandable. In addition, this method is more appropriate for handling possibly nonstationary or shorter time series than is the frequency domain approach. However, in order to obtain the valid results, it is necessary to have enough data points in a time perspective.

These analyses work well as long as the assumptions are satisfied. The number of time points, t, should be big enough for the validity of the model. For spectral analysis, it is necessary to obtain stationarity. However, if one has a series with as small as 30 or fewer time points, then both approaches don't have consistent results. Other than that, if one has unlabeled data, which means that one has no prior information about the data, then one is not sure whether the data is stationary or nonstationary. In that case, it is hard to tell which method would be better to explain the given time series data.

## 1.5 Time series clustering

Like static data, it is sometimes desirable to determine groups of similar time series when the time series data are unlabeled. For example, in the stock market, one could identify companies whose stock prices have similar patterns. Or, one would like to determine products with similar selling patterns for a certain time period or to group sales regions with similar pattern for the time being.

To this end, the definition of similarity of two time series is considered first. Given a set of time series sequences, $X = x_1, x_2, ..., x_n$ and $Y = y_1, y_2, ..., y_n$, the similarity of two series is expressed as $Sim(X, Y)$. Even though there is no consensus of definition of time series similarity, it could be stated that two time series are similar if they have enough non-overlapping time ordered subsequences that are similar. The two subsequences are considered to be similar if one is enclosed within an envelope of a user defined width around another (Negi and Bansal 2005). That is, the similarity measure does not find exactly matching subsequences but allows for imprecise matches. However, the similarity search algorithm should be efficient enough to find similar time series as fast as possible. Thus, one should consider precision and efficiency

simultaneously when one develops an algorithm. After development, the algorithm can be used for indexing the time series, clustering, rule discovery, and so on.

There have been various algorithms developed to cluster different types of time series data. According to Liao's categorization, there are two major groups of them: modifying the existing algorithms for clustering static data for time series data, the so-called 'raw-data-based approach', and converting time series data into the form of static data such as modeling the data or extracting specific features from the time series so that the existing algorithms for clustering static data can be directly used, so-called 'feature-extraction or model-based approach' (Liao 2005).

As mentioned before, the appropriate method of time series clustering depends on the characteristics of the data. Time series data can be distinguished in several perspectives, such as whether the data are discrete-valued or real-valued, uniformly or un-uniformly sampled, univariate or multivariate, stationary or nonstationary, and whether the series are of equal or unequal length. Un-uniformly sampled data can be converted into uniformly sampled data. There has been a wide range of research to cluster univariate time series with different data characteristics. We will discuss this in the next chapter.

Many time series arising in practice are considered as components of some multivariate time series $\{X_t\}$ whose specification includes not only the serial correlation of each component series $\{X_{ti}\}$ but also the cross-correlation between different component series $\{X_{ti}\}$ and $\{X_{tj}\}$. Thus, ignoring or underestimating the dependences between the variables can lead to serious errors in measurement.

## 1.6 Proposition

In this paper, a new clustering methodology for multivariate time series data is proposed. We assume that a time series consists of a sequence of real numbers which represent the values of a measured parameter at equal intervals of time. Also, it is assumed that an observation contains multiple numbers of time series data which are cross-correlated. For example, one would like to know whether the various indicators of a patient's health measured over time are being produced by a patient who is likely to live or one that is likely to die. For another example, a credit scoring company would want to know whether its customers will turn out to be good, bad, or default customers, although data have been recorded for only a short time period.

There has been a lot of research on clustering time series data in various areas: the stock market, regional sales in marketing, health management, image reconstruction, robot movement study, and so on. However, it is surprising that there are not many studies dealing with multivariate time series data. Even if a study considered multiple variables, many of them ignored the correlation between the variables. The primary reason for this would be the complication of high dimensionality of the multivariate time series data.

A natural approach to deal with this high dimensionality difficulty is dimension reduction. This includes indexing time series or extracting selected features. The key to the procedure is extracting a few key "features" for each time series. That is, the objective is to map each time sequence $X$ to a point $f(X)$ in the relatively low dimensional feature space. There have been several methods developed for dimension reduction: Discrete Fourier Transform (DFT), Discrete Wavelet Transform (DWT), Singular Value Decomposition (SVD) (Korn, Jagadish and Faloutsos 1997), Principal Component Analysis (PCA), and so on. Among them, DFT showed that the method can deal with multivariate time series data (Kakizawa, Shumway and Taniguchi

1998). But, as mentioned before, this method relies on the coefficients of a spectral density model; there must be a large enough number of temporal data points for model validity. This method is well suited for naturally occurring signals, which are sinusoidal in nature, but is ill suited for others.

Principal component analysis is one of the most popular feature extraction methods. Original principal components are linear functions computed on a single two-mode, for example, objects × variables, matrix (Flury 1988). For the multivariate time series data, the two-mode constraint should be relaxed to k-mode, such as objects × variables × time. Since this method was introduced by Kapteyn et al. in 1986, the multi-way principal component method has been used successfully, especially for statistical process control (Kosanovich, Piovoso, Dahl, MacGregor and Nomikos 1994).

Since both PCA and MPCA are linear algorithms, they could be beyond the capabilities for the nonlinear structure like multivariate time series data. Kernel variant Principal Component Analysis (KPCA) is a nonlinear version of PCA (Sholkopf and Smola 2002). In machine learning area, kernel-based methods have been intensively used in the last decade. The gist of kernel methods is mapping the data implicitly into a high-dimensional feature space without explicitly using or even knowing the mapping. Therefore, this allows computing scalar products in spaces, where one could otherwise hardly perform any computations (Muller, Mika, Ratsch, Tsuda and Scholkpf 2001).

We introduce a novel methodology for multivariate time series clustering based on the Multi-way Principal Component Analysis (MPCA) with kernel variant and distance measurement methods of latent variables. After that, we attempt to show the guidelines to

determine for which states the introduced method is effectively working well based on various cases of simulation. Field data test results will follow after the simulation results.

In Chapter 2, we review methods of univariate and multivariate time series clustering research. In Chapter 3, original Principal Component Analysis and the Multi-way Principal Component Analysis are briefly explained. After that, Kernel Multi-way Principal Component Analysis is presented and compared with the preceding methods. In Chapter 4, the procedures including any assumptions and concerns relating to the simulation are discussed. Various cases of multivariate time series data generation are discussed in Chapter 4 as well. Chapter 5 brings us to the results of the simulation study. In Chapter 6, we show an example of real data and see if the method introduced here is working effectively. In Chapter 7, we conclude this study with the discussion of study limitations and the direction of possible future research.

# CHAPTER 2

# LITERATURE REVIEW

Time series data analysis has a long history and has been studied extensively in various fields, including economics, statistics, process control theory, signal processing, and so on. However, surprisingly, there is limited research on time series clustering. Most of the time series clustering studies are from the late 1980s. Recently, with sizable datasets available thanks to the advanced technology of computation and data storing devices, this field has had another look from the practitioners and field operators, especially in computer science and data mining. During the last decade, time series clustering has been popular, and a lot of research has contributed to suggest new methods for various time series data contexts.

In this chapter, we review several research articles which suggested specific methods of time series clustering for their own context. Although there is considerable literature available concerning clustering methodologies and applications, only a few applications have been reported that cluster multivariate time-series data since time series data is complicated in higher dimensions. So, we will especially pay attention to the methods which attempted to cluster multivariate time series.

In addition, due to the various and complex situations, such as different time lengths, missing values, discrete-valued or real-valued, and so on, the algorithms that showed good performance cannot be applied to other contexts. In other words, it is very hard to compare

suggested algorithms with each other due to their various environments. Therefore, this review will introduce several studies which worked well for each situation and look through the pros and cons.

## 2.1 Three approaches to time series clustering

According to several survey research articles so far (Liao 2005), it can be said that there are three major categories of time series clustering algorithms: raw-data-based, feature-extraction-based, and model-based approaches.

### 2.1.1 Raw-data-based approach

Raw-data-based approaches are methods that work with raw data. The two time series being compared are sampled at the same interval, but their length or the number of time points might or might not be the same. For each time point, one seeks to find a group which is close in the perspective of distance. In the end, the time series which are close enough for each time point are considered as the same group. To measure the closeness, various distance measures are used such as Euclidean distance, the root mean square distance, statistical distance, or Minkowski distance.

However, handling raw data directly implies dealing with high dimensional space. When there are only a small number of time points, this method would be appropriate since it is quick, simple, and easy to perform. But, it has limitations for time series with a large number of temporal points. In addition, if there are more than one variable, and the variables have cross-correlation; this process requires extra steps to do so. For clustering multivariate time varying data, Kosmelj and Batagelj (1990) modified the relocation clustering procedure that was originally developed for static data. For measuring the dissimilarity between trajectories as required by the procedure, they first introduced a cross-sectional approach-based general model

that incorporated the time dimension, and then developed a specific model based on the compound interest idea to determine the time-dependent linear weights. However, the proposed cross-sectional procedure ignores the correlations between variables over time and works only with time series of equal length. To form a specified number of clusters, the best clustering method among all the possible methods is the one with the minimum generalized Ward criterion function. Also, it would be difficult to work with highly noisy data.

### 2.1.2 Model-based approach

To solve the problems of the raw-data-based approach, the feature extraction and model-based approaches are suggested. Even though it is possible to avoid working in the high dimensional space using feature-extraction methods, a specific feature extraction method works in a certain context or data set. Therefore, it is more generalized to have a time series model, which is model-based approach. The model-based approach considers that a time series is generated by a certain type of model or by a mixture of underlying probability distributions. Time series are considered similar when the models characterizing individual series or the remaining residuals are similar after model fitting. The Box-Jenkins model has often been used for both univariate and multivariate cases.

Baragona (2001) used ten sets of different ARMA models and vector ARMA (VARMA) models for a simulation test. The author compared three meta-heuristic methods for partitioning a set of time series into several clusters with the criterion of cross-correlation between the time series. For this, he suggested the cross-correlation maximum absolute value between each pair of time series that belongs to the same cluster. The cross-correlation is computed from the residuals of the models of the original time series. In this study, he regarded time series similarity as cross-correlation only. So, if the correlation between the time series is not strong enough, it did not

show effective clustering results. Also, this method works well only with Box-Jenkins time series model since it defines the similarity as the extent of correlation.

Maharaj (2000) also worked with Box-Jenkins model time series data. She developed an agglomerative hierarchical clustering procedure that is based on the *p*-value of a test of hypothesis applied to every pair of given stationary time series. The author assumed that each stationary time series can be fitted by a linear AR(k), where k is the order of auto-regressive. So, a model can be denoted by a vector of parameters. After that, a $\chi^2$ test statistic was derived to test the null hypothesis that there is no difference between two stationary time series. Two series were grouped together if the associated *p*-value was greater than the pre-specified significance level. Beside the fact that this method worked with only univariate time series data, there were some cons. If the size of the time series was large, comparing one by one could be impossible. And, if there is cross-correlation between the time series, the test results would mask the effects.

One approach for this model-based class is using a Markov model or a Hidden Markov model. Ramoni et al. (Ramoni, Sebastiani and Cohen 2001) presented a Bayesian algorithm for clustering by dynamics (BCD). Given a set *S* of n numbers of univariate discrete-valued time series, BCD transforms each series into a Markov chain (MC) and then clusters similar MCs to discover the most probable set of processes. Considering a partition as a hidden discrete variable *C*, each stat $C_k$ of *C* represents a cluster of time series, and hence determines a transition matrix. The task of clustering is regarded as a Bayesian model selection problem with the objective to select the model with the maximum posterior probability. The similarity between two estimated transition matrices is measured as an average of the symmetrized Kullback-Leibler distance between corresponding rows in the matrices. The objective is to find a maximum posterior

probability partition of a set of MCs. This method is considering only single and discrete-valued time series data.

Also the authors (Ramoni, Sebastiani and Cohen 2000) showed the same transition probabilities model for robot sensor data which has multiple real-valued variables. They transformed the data into discrete-valued data as for univariate cases. However, they ignored the cross-correlation between the variables. In addition, this method cannot be applied to other real-valued data such as stock price data or patients' physical data since it too approximated the data with discrete values for the procedure.

Considering that a set of multivariate, real-valued time series is generated according to hidden Markov models, Oates et al. (Oates, Firoiu and Paul 1999) presented a hybrid clustering method for automatically determining the $k$ number of generating Hidden Markov Models (HMMs), and for learning the parameters of those HMMs. They used the same robot sensor data as above. HMMs are statistical models of sequential data that have been used successfully in many machine learning applications. This assumes that the observation at time $t$ was generated by some process whose state $S_t$ is hidden from the observer and that the state of this hidden process satisfies the Markov property: given the value of $S_{t-1}$, the current state $S_t$ is independent of all the states prior to $t-1$. Also, they used the discrete HMMs; thus it would not be easy to apply this method to the data we are focusing on in this study.

Another approach that considers multivariate time series is the Discrete Fourier Transform (DFT) algorithm. This method belongs to frequency domain time series analysis. In mathematics, the Fourier transform is a certain linear operator that maps functions to other functions (Shumway 1988). Loosely speaking, the Fourier transform decomposes a function into a continuous spectrum of its frequency components, and the inverse transform synthesizes a

function from its spectrum of frequency components. Agrawal et al. (Agrawal, Faloutsos and Swami 1993) used DFT for simple univariate time series data to approximate the single time series. After transformation, they argued that only the first *k* coefficients are required for representing the time series.

Even though this method is working well for specific cases, including sinusoidal data, seismology data, sound data, etc., most other time series data don't work with the method. If the data has a complex structure so that the first *k* coefficients are large, then the efficiency of this method drops drastically. But, much research has been conducted using DFT with the Kullback-Leibler (*KL*) discrimination information measure. To group multivariate vector series of earthquakes and mining explosions, Kakizawa et al. (Kakizawa, Shumway and Taniguchi 1998) applied hierararchical clustering as well as k-means clustering. They measured the disparity between spectral matrices corresponding to the $X_t$ matrices of autocovariance functions of two zero-mean vector stationary time series with two quasi-distances: the J divergence and the symmetric Chernoff information divergence which are other versions of KL distance measure. Shumway (2003) investigated the clustering of nonstationary time series by applying locally stationary versions of KL discrimination information measures that give optimal time-frequency statistics for measuring the discrepancy between two non-stationary time series. To distinguish earthquakes from explosions, an agglomerative hierarchical cluster analysis was performed until a final set of two clusters was obtained.

In sum, model-based approaches can obtain more generalized results than any other methods. Using the time domain or frequency domain approach, they can deal with multivariate cases, for example, VARMA and spectral density estimation. Also, these models can handle large data sets with many temporal points and a large number of observations. But, this could be

a con because without a large number of observations and big enough temporal points, the model validity is not very good. Therefore, for this study, the model-based approach is not considered.

### 2.1.3 Feature-extraction approach

The difficulty of time series clustering comes mostly from the high dimensionality of the data, which has many temporal points as well as multiple variables to consider simultaneously. To solve the concerns of raw-data methods, which are not effective in working with high dimensional time series, several feature-extraction methods have been proposed. Therefore, dimension reduction has been a natural approach for clustering in time series data especially in the multivariate case. For the computer science field, it is called "time series indexing" (Agrawal, Lin, Sawhney and Shim 1995). Most of the approaches for performing clustering in time series data developed so far rely on dimension reduction (Agrawal, et al. 1993).

The main idea of this method is to identify a certain feature from a time series which represents the time series well with, hopefully, few feature values. That means, even if the temporal data is measured at many different time points, a few underlying temporal points may account for much of the data variation. Of course, this may lead to some loss of information. So, the objective of time series clustering using indexing is to find the most efficient way of high dimensional time series data clustering without losing much important information.

There are several ways of performing dimension reduction on time series data: Discrete Fourier Transform (DFT), Discrete Wavelet Transform (DWT), Piecewise Aggregate Approximation (PAA), and Singular Value Decomposition (SVD) (Toshniwal and Joshi 2005).

The basic algorithm of these methods is to extract a few key features such as scores or latent variables for singular value decomposition: spectral matrices for the discrete Fourier transform, wavelets for the discrete wavelet transform, and so on. Here, it is interesting that

researchers consider DFT differently. Liao placed DFT into the raw-data-based approach because this method dealt with the frequency domain time series directly. Several data-mining researchers considered this method a feature extraction approach, however, because it transforms time-domain time series into frequency-domain data so that the time series can be explained as a few Fourier functions. Also, this could be considered a model-based method by researchers in economics. In sum, the algorithm of feature extraction is that one maps each time sequence *X* to a point *f(X)* in the relatively low dimensional feature space such that the similarity between the time series *X* and *Y* is approximately equal to the Euclidean distance between the two points *f(X)* and *f(Y)*.

The DFT approach considers a time series in the frequency domain so that a given time signal is transformed to obtain a set of Fourier coefficients. After that, the specific variances for each time series data set, the so-called spectra, are obtained. In many practical signals, the most information of the signals is concentrated in the first few Fourier coefficients. But, to obtain valid spectra from the data set requires a huge data set. In addition to that, this method is appropriate for naturally occurring signals such as sounds, seismology data, and so on.

For the wavelet transform, the Haar wavelet transform is primarily used (Chan and Fu 1999). The basis function for Haar is not smooth and as a result the Haar wavelet transform approximates any time series by a ladder-like structure. Therefore, this transform still needs many coefficients to approximate a smooth function. So the number of coefficients to be added must be high. In case of PAA, the time sequence is divided into equal length segments. The corresponding feature sequence comprises mean values of each segment. But the means representing each segment give only a rough approximation of each time sequence. And, both methods, DWT and PAA, cannot handle multiple time series simultaneously. Thus, if multiple

variables are considered, these methods need another solution to deal with multivariate time series together.

In the next chapter, we address one of the feature extraction methods and pros and cons. After that, we introduce an adjusted method of feature extraction with which we can cluster multivariate time series with complex structure and cross-correlation.

# CHAPTER 3

# METHODS FOR MULTIVARIATE TIME SERIES CLUSTERING

Principal Component Analysis (PCA) is one of the popular feature-extraction methods. If PCA is applied to time series data, it can provide an efficient way to find the underlying temporal characteristics and reduce the input dimensions. This linear transformation has been widely used in various fields. If the data are concentrated in a linear subspace, PCA provides a way to compress data and simplify the representation without losing much information. However, if the data are concentrated in a nonlinear subspace, PCA will fail to work well. In this case, one may consider kernel principal component analysis (KPCA) for time series data since the data would exhibit nonlinearity due to the many interactions between variables and auto-correlation of temporal time points. KPCA is a nonlinear version of PCA (Sholkopf and Smola 2002) that has been studied intensively in the last decade.

For the multivariate time series data, we can borrow *k*-mode PCA method (Kapteyn, Neudecker and Wansbeek 1986) which removes the two-mode restriction from the original PCA assumptions. Multi-way Principal Component Analysis (MPCA) has been tried and applied successfully to characterize the data matrices since 1980s in the quality control analysis. Using MPCA with kernel variant is the ultimate destination of this research. In this chapter, a novel clustering method for multivariate time series, MPCA with kernel variant will be introduced and explained.

## 3.1 Data matrix

A time series dataset with N observations and T temporal points can be conveniently represented by the following matrix:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1T} \\ x_{21} & x_{22} & \cdots & x_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NT} \end{bmatrix} \tag{3.1}$$

where $x_{ij}$ is the measurement of the expression level of $i^{th}$ time in sample j.

In this paper, multivariate time series with the same time length and real-valued variables are considered. Each observation has P variables which have T temporal points respectively. This could be represented as follows:

$$X_1 = \begin{bmatrix} x_{111} & x_{112} & \cdots & x_{11T} \\ x_{121} & x_{122} & \cdots & x_{12T} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1P1} & x_{1P2} & \cdots & x_{1PT} \end{bmatrix}, \ldots, X_N = \begin{bmatrix} x_{N11} & x_{N12} & \cdots & x_{N1T} \\ x_{N21} & x_{N22} & \cdots & x_{N2T} \\ \vdots & \vdots & \ddots & \vdots \\ x_{NP1} & x_{NP2} & \cdots & x_{NPT} \end{bmatrix}$$

$$\text{or } X_i = \begin{bmatrix} x_{i11} & x_{i12} & \cdots & x_{i1T} \\ x_{i21} & x_{i22} & \cdots & x_{i2T} \\ \vdots & \vdots & \ddots & \vdots \\ x_{iP1} & x_{iP2} & \cdots & x_{iPT} \end{bmatrix} \text{ where } i = 1, 2, \ldots, N. \tag{3.2}$$

## 3.2 Principal Component Analysis for time series data

PCA provides an approximation of a data matrix, *X*, in terms of the product of two small matrices, *P* and *Q*. The *P* matrix usually explains the dominant "object pattern" of the information, and the *Q* matrix shows the complementary "variable pattern". (Wold, Esbensen and Geladi 1987a) For example, a data matrix could represent *N* potential patients and *T* time points. Once the groups of the patients are obtained by using PCA, then each group corresponds to an object pattern and each object pattern would also be explained by a variable pattern.

Generally, a variance-covariance matrix is used for the analysis instead of using the *X* data matrix since the variance-covariance matrix has lower dimension and contains most of the original data information.

PCA provides us with a set of orthogonal axes along which we can project the data, hopefully allowing us to account for most of the data with just the first few axes in the new space. In other words, it efficiently represents the data by finding orthogonal axes which are maximally not correlated with the data. PCA assumes normality and the data should be independent and identically distributed. Therefore, PCA is an appropriate model for data that are generated by a Gaussian distribution, or data that are best described by second-order correlations. That is, if the data is suspected of having higher order correlation, then the results can not be trusted. In addition, the results could be affected sensitively by extreme data points.

Singular value decomposition (SVD) is another name of PCA used in numerical analysis (Wold, et al. 1987a). Korn et al. used SVD and SVD with delta (SVDD) for dimension reduction of huge collections of time sequences, where delta is the difference between the actual value and the value that SVD reconstructs (Korn, Jagadish and Faloutsos 1997). They considered univariate time queries though; their results showed that PCA is one of the better methods to index time series data.

For the standard PCA algorithm, given a set of centered observations $x_i \in \Re^N$, where $i$ = 1, 2, …, $N$ and $\sum_{i=1}^{N} x_i = 0$, PCA diagonalizes the covariance matrix

$$C = \frac{1}{N} \sum_{i=1}^{N} x_i x_i^T \ . \tag{3.3}$$

To do this, one has to solve the eigenvalue equation $\lambda v = Cv$ for eigenvalues $\lambda \geq 0$ and eigenvectors $v \in \Re^N$.

As $\lambda v = Cv = {}^{1}\!/_{N} \sum_{i=1}^{N} (x_i \cdot v) x_i$, all solutions *v* with $\lambda \neq 0$ must lie in the span of $x_1, x_2, \cdots, x_N$; hence we can put $x_i$ into both sides of $\lambda v = Cv$. Then, we have $\lambda(x_i \cdot v) = (x_i \cdot Cv)$ for all *i* = 1, …, *N*. After obtaining, hopefully, a few eigenvalues and eigenvectors to explain most of the information, we can get the principal components or scores from $\hat{y}_i = \hat{e}_i^{'}(x - \bar{x})$ where *i* = 1, 2, …, *N*.

3.3 Kernel method

If we want to analyze data with high order interactions or nonlinearity, Kernel variant PCA (KPCA) would be one of the alternatives. In here, kernel is the main concept. Before explaining KPCA, it is important to first understand the concept of kernel in machine learning. Kernel methods for machine learning were introduced by Nadaraya and Watson in 1964 (Friedman 2006). After that, even though there are several disadvantages of the method, this has been applied to various solutions. It starts with a basic learning theory problem.

Suppose we have two classes of objects. Now we have to assign a new object to one of two groups, such as: $(x_1, y_1), \cdots, (x_n, y_n) \in \aleph \times \{\pm 1\}$.

For this task done, we have to have a ruler to measure which one belongs to which group. And, the choice of the similarity measure of the inputs is a major question especially in the machine learning. So, let's consider a similarity measure of the form

$$\begin{aligned} &k : X \times X \to \Re \\ &(x, x') \mapsto k(x, x'), \end{aligned} \tag{3.4}$$

that is, a function that, given two patterns $x$ and $x'$, returns a real number characterizing their similarity. The function *k* is called a kernel (Sholkopf, et al. 2002). To find a general similarity measures is difficult. So, we can apply a simple function for this problem, a dot product. A dot

product is defined as $\langle x, x' \rangle := \sum_{i=1}^{N} \underline{x}_i \underline{x}_i'$, where $\underline{x}_i$ and $\underline{x}_i'$ are $i^{\text{th}}$ vector of $x$ and $x'$. Using the inner product as a kernel function can have other benefits: (1) it computes the cosine of the angle between the vectors $\underline{x}$ and $\underline{x}'$, if they are normalized to length 1, (2) it enables us to compute the length of a vector as $\|x\| = \sqrt{\langle x, x \rangle}$, and thus (3) the distance between two vectors can be computed as the length of the difference vector. That is, if we can use the inner product or if we have data vectors in the inner product space, the similarity of the vectors can be computed using vector distance measures. Therefore, we can transform the data into a feature space which enables us to use the inner product so that we can reap the benefits of it.

However, a question remains how kernel method can work with nonlinearity. Suppose we have a predictor space which cannot be classified with a simple, linear decision rule. Then, as stated above, it might be reasonable to use a mapping that we can transform the original data into a new space which can deal with a potentially much higher dimensional space such as feature space *F*:

$$\begin{aligned} &\Phi : \Re^N \to F \\ &x \mapsto \Phi(x) \end{aligned} \tag{3.5}$$

This transformation invokes another question, the curse of dimensionality, which is the difficulty that an estimation problem increases drastically with the dimension *N* of the space, since as a function of *N,* one needs exponentially many patterns to sample the space properly.

However, the contrary can be true: learning in *F* can be simpler if one uses low complexity, simple linear classifiers. Muller et al. explained this well using a toy example (Muller, Mika, Ratsch, Tsuda and Scholkpf 2001).

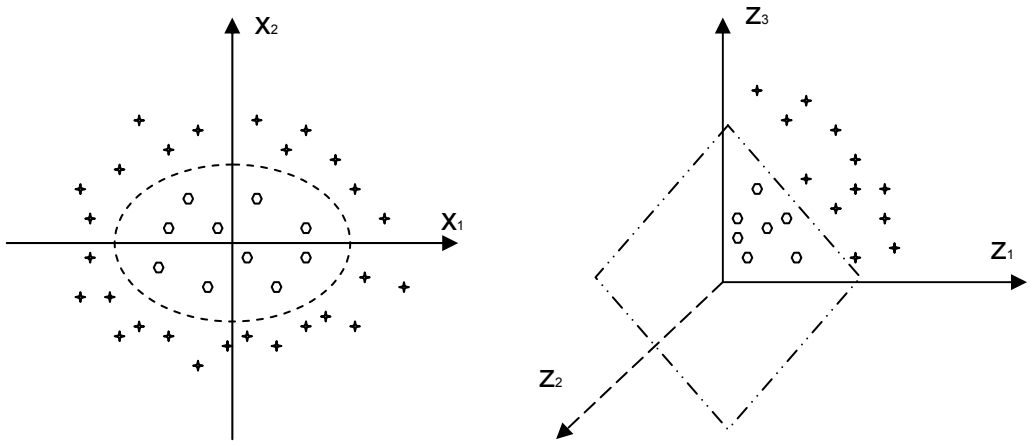


Figure 3.1. Two-dimensional Classification Examples

The examples in Figure 3.1 show that the input space has two dimensions to be classified with a circular classifier. Then, if we map the predictor space to a feature space of second-order monomials:

$$\begin{aligned} &\Phi : \Re^2 \to \Re^3 \\ &(\underline{x}_1, \underline{x}_2) \mapsto (\underline{z}_1, \underline{z}_2, \underline{z}_3) := \left(\underline{x}_1^2, \underline{x}_1\underline{x}_2, \underline{x}_2\underline{x}_1, \underline{x}_2^2\right) \end{aligned} \tag{3.6}$$

a linear hyperplane is enough for a classifier. Of course, the complexity is the greatest concern for large dimensional real-world problems.

However, for certain feature spaces *F* and corresponding mappings $\Phi$ there is a highly effective trick for computing inner products in this high dimensional feature spaces without explicitly mapping into the spaces by means of kernels nonlinear in the input space (Muller, et al. 2001). Like before, we use kernel representations of the form

$$k(x, x') = \left\langle \Phi(x), \Phi(x') \right\rangle, \tag{3.7}$$

which enables us to compute the value of the dot product in feature space without having to explicitly compute the map $\Phi$. Now, we use the same example above.

For the map, we have $\Phi : (\underline{x}_1, \underline{x}_2) \mapsto \left(\underline{x}_1^2, \underline{x}_2^2, \underline{x}_1\underline{x}_2, \underline{x}_2\underline{x}_1\right)$.

Then,

$$\left\langle \Phi(x), \Phi(x') \right\rangle = \left( \underline{x}_1^2, \underline{x}_1 \underline{x}_2, \underline{x}_2 \underline{x}_1, \underline{x}_2^2 \right) \begin{pmatrix} \underline{x}'^2_1 \\ \underline{x}'_1 \underline{x}'_2 \\ \underline{x}_2 \underline{x}_1 \\ \underline{x}'^2_2 \end{pmatrix} = \underline{x}_1^2 \underline{x}'^2_1 + \underline{x}_2^2 \underline{x}'^2_2 + 2\underline{x}_1 \underline{x}_2 \underline{x}'_1 \underline{x}'_2 = \left( \underline{x}_1 \underline{x}'_1 + \underline{x}_2 \underline{x}'_2 \right)^2 \tag{3.8}$$

$$= \left( (\underline{x}_1, \underline{x}_2) \left( \underline{x}'_1, \underline{x}'_2 \right)^T \right)^2 = (\underline{x} \cdot \underline{y})^2$$

The inner product of two vectors in a feature space can be readily reformulated in terms of a kernel function *k*. The key point here is that the inner product can be implicitly computed in a feature space without explicitly using or even knowing the mapping, which is seldom computable. This result could be expanded; any procedure which uses inner products only could have a nonlinear version of this algorithm.

To summarize, once we map the data into a feature space using an inner product kernel function, we can define a similarity measure from the dot product in the feature space. Also, it allows us to deal with the patterns geometrically. Finally, even if the data sets have high order interactions between the variables or nonlinearity, the mapping turns the data into a nonlinear feature space so that we can enjoy the benefits of the inner product in the feature space without explicitly computing complex calculations.

## 3.4 PCA with Kernel method (KPCA)

Now, we can apply a kernel function to PCA since PCA uses an inner product for the procedure. Instead of using a predictor variable space, we can map input space into an arbitrarily large, possibly infinite dimensional feature space *F*.

$$\begin{aligned} &\Phi : \Re^N \to F \\ &x \mapsto \Phi(x) \end{aligned} \tag{3.9}$$

Thus, we work with the sample $(\Phi(x_1), y_1), (\Phi(x_2), y_2), \cdots, (\Phi(x_N), y_N) \in F \times Y$ instead of the sample data $x_1, x_2, \cdots, x_N \in \Re^N$.

To derive kernel PCA, we first map the data $x_1, x_2, \cdots, x_N \in \Re^N$ into a feature space $F$ and compute the covariance matrix

$$C^* = \frac{1}{N}\sum_{j=1}^{N} \Phi(x_j)\Phi(x_j)^T . \tag{3.10}$$

The principal components are then computed by solving the eigenvalue problem: find

$$\lambda > 0, V \neq 0 \text{ with } \lambda V = CV = \frac{1}{N}\sum_{j=1}^{N} (\Phi(x_j) \cdot V)\Phi(x_j) . \tag{3.11}$$

As in the original PCA above, all solutions $V$ with $\lambda \neq 0$ lie in the span of

$$\Phi(x_1), \Phi(x_2), ..., \Phi(x_N) . \text{ So, } V \text{ can be written as } V = \sum_{j=1}^{N} \alpha_i \Phi(x_j). \tag{3.12}$$

By multiplying with $\Phi(x_k)$ from the left, $\lambda V = CV$ would be

$$\lambda(\Phi(x_k) \cdot V) = (\Phi(x_k) \cdot CV) \text{ for all } k = 1, \ldots, N. \tag{3.13}$$

Now, the formula (3.11) can be expressed

$$\lambda \sum_{j=1}^{N} \alpha_i (\Phi(x_k) \cdot \Phi(x_j)) = \frac{1}{N}\sum_{j=1}^{N} \alpha_j \left( \Phi(x_k) \cdot \sum_{i=1}^{N} \Phi(x_i)(\Phi(x_i) \cdot \Phi(x_j)) \right) \text{ for all } k = 1, \ldots, N. \tag{3.14}$$

Defining an $N \times N$ matrix $K$ by $K_{ij} \approx (\Phi(x_i) \cdot \Phi(x_j))$, this leads $N\lambda K\alpha = K^2\alpha$, where $\alpha$ denotes the column vector with entries $\alpha_1, \ldots, \alpha_N$. To find solutions of this, one solves the eigenvalue problem $N\lambda\alpha = K\alpha$ for nonzero eigenvalues.

The solutions $(\lambda_k, \alpha^k)$ further need to be normalized by imposing $\lambda_k(\alpha^k \cdot \alpha^k) = 1$ in *F*. Also, the data needs to be centered in *F*. This can be done by simply substituting the kernel matrix *K* with $\hat{K} = K - 1_N K - K 1_N + 1_N K 1_N$ where $(1_N)_{ij} = \frac{1}{N}$. (3.15)

For extracting features of a new pattern *x* with kernel PCA, one simply projects the mapped pattern $\Phi(x)$ onto the eigenvector $V^k$

$$\left(V^k \cdot \Phi(x)\right) = \sum_{i=1}^{N} \alpha_i^k \left(\Phi(x_i) \cdot \Phi(x)\right) = \sum_{i=1}^{N} \alpha_i^k k(x_i, x). \quad (3.16)$$

The difference between original PCA and KPCA is that one has to deal with the number of sample dimension, $n \times n$, instead of the number of variables, $p \times p$.

In summary, the following steps are necessary to compute the principal components: first, compute the matrix *K*; second, compute its eigenvectors and normalize them in *F*; third, compute projections of a test point onto the eigenvectors which are scores, principal components, or latent variables to be clustered.

## 3.5 Kernel variant Multi-way Principal Component Analysis (KMPCA)

For multivariate time series data, one could face several problems:

(1) The dimension is usually very large. In reality, there are a lot of variables to handle. (2) And, they are more or less correlated with each other. Like interaction effects in the experimental design field, there are intrinsic effects between the variables. (3) In addition, these days it is common to deal with a sizable number of observations in a short amount of time. So, it is necessary to handle as much data as possible at once from the perspective of cost and time.

Multi-way Principal Component Analysis (MPCA) is an extension of PCA to handle data in three dimensional arrays (Kosanovich, Piovoso, Dahl, MacGregor and Nomikos 1994). Since Kapteyn et al. introduced the k-mode PCA in 1986 by relaxing the two-mode restriction, many

researchers have applied this method to various situations during last decade (Kapteyn, et al. 1986). Wold et al. introduced this method with a name of multi-way PCA for simulation data and applied it to an experiment of multivariate calibration with liquid chromatography and a UV array detector (Wold, Geladi, Esbensen and Ohman 1987b). Nomikos and MacGregor (Kosanovich, et al. 1994) showed that MPCA is suitable for handling multivariate batch data in practice.

The procedure is as follows: first, one should unfold the three-dimensional array $X_i$ (objects × variables × time) as above slice by slice and rearrange the slices into a large two-dimensional matrix $X$ (objects × (variables·time)). After that, a regular PCA is performed. Consider the $N \times PT$ matrix, $X$**,** given by

$$X = \begin{bmatrix} x_{111} & x_{121} & \cdots & x_{1P1} & x_{112} & x_{122} & \cdots & x_{1P2} & \cdots\cdots & x_{11T} & x_{12T} & \cdots & x_{1PT} \\ x_{211} & x_{221} & \cdots & x_{2P1} & x_{212} & x_{222} & \cdots & x_{2P2} & \cdots\cdots & x_{21T} & x_{22T} & \cdots & x_{2PT} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{N11} & x_{N21} & \cdots & x_{NP1} & x_{N12} & x_{N22} & \cdots & x_{NP2} & \cdots\cdots & x_{N1T} & x_{N2T} & \cdots & x_{NPT} \end{bmatrix}. \qquad (3.17)$$

This matrix consists of rows, $n$ = 1, 2, ..., $N$, displaying measurements for variables $X_p$, $p$ = 1, 2, …, $P$, at time period $t$, $t$ = 1, 2, …, $T$. Each row contains the complete set of data collected for object $i$, $i$ = 1, 2, …, $N$. Consider the first row of $X$. The values of $x_{111}$ through $x_{1PT}$ represent all observations collected for object $i$ = 1. Values $x_{111}$ through $x_{1P1}$ represent the first observation collected for each of the $P$ variables. The values $x_{11T}$ through $x_{1PT}$ represent the final values collected for each of the $P$ variables with respect to the first object.

The results after the transformation and the analysis are interpreted the same as in ordinary PCA. The scores or principal components obtained from the analysis can explain most information contained in the data. However, MPCA is a linear algorithm as is PCA. So, if one

has data with higher order correlations and nonlinearity, one should consider other treatments for more reliable results.

For a multivariate time series data, one can use kernel variant MPCA. In MPCA, one can unfold multivariate time series data cube so that one can apply ordinary PCA to the data as for univariate time series data. KMPCA is nothing but the same MPCA with kernel method. So, the procedure will be: (1) first, unfold a three-mode data cube into a two-mode data matrix, (2) secondly, obtain the kernel matrix, $\boldsymbol{K}$, from the two-mode data matrix, and (3) lastly, by using kernel matrix, perform the ordinary PCA. After the procedure, one can get, hopefully, a few latent variables which can explain most of the data information. Based on the score values of each object, we can cluster them with the static clustering methods mentioned in Chapter 1.

# CHAPTER 4

# METHODOLOGY FOR EVALUATING KMPCA

In this chapter, the methods for investigating clustering performance of Multi-way Principal Component Analysis with kernel variant (KMPCA) are explained. A simulation study will show if the proposed method would work on clustering multivariate time series data and provide the guidelines for selecting appropriate clustering schemes to apply KMPCA to.

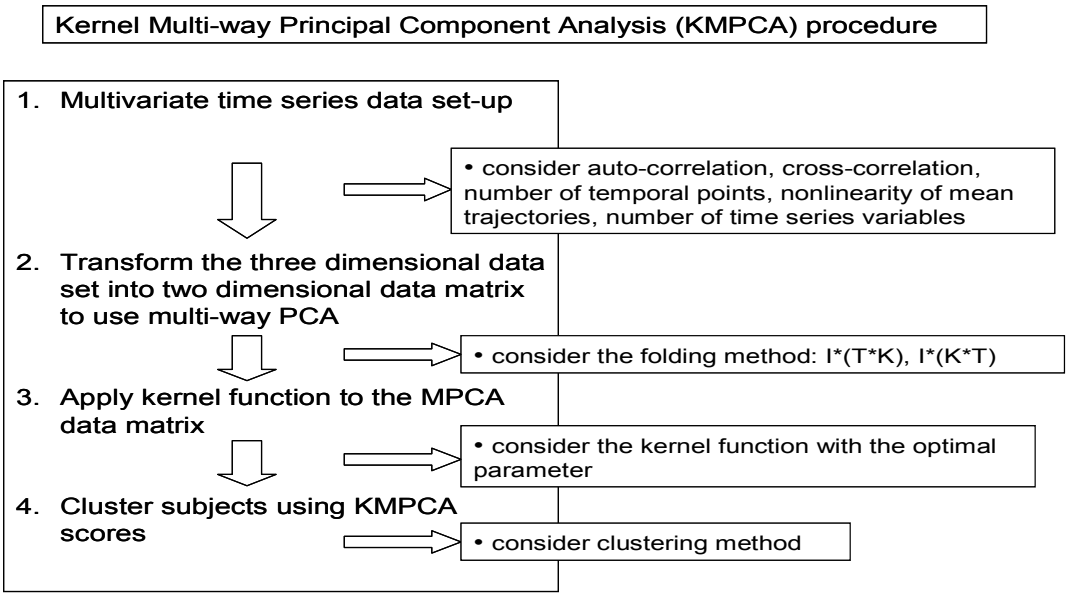


Figure 4.1. Kernel Variant Multi-Way Principal Component Analysis (KMPCA) Procedure Flow Chart

KMPCA procedure has multiple stages to administrate as seen in figure 4.1. Therefore, it is necessary to consider these steps one after another to generate simulation data which contains all the factors and cases we are interested in for this study.

To generate non-linear multivariate time series data, we used non-linear function to generate mean profiles for different groups. Multivariate time series data has multi dimensional data structure since there are associations among the time points, various levels of cross-correlation and auto-correlation. Also, we have to consider how much the error would be dispersed around the mean profiles as well. If the number of temporal points increases, a time series can be higher degree of polynomial function and affect the clustering performance. Therefore, number of temporal points should be investigated. However, the number of groups to cluster and the number of time-series variables are not considered in this study. We generated tri-variate time series and set up two groups for the simulation study.

After generating the data, the data should be transformed from three-dimensional data matrix to two-dimensional, folded data matrix for applying MPCA to. Three dimensions are observation ($I$), time ($T$), and variables ($P$). Therefore, there could be 6 ($= {}_3C_2$) different variations of folding. However, we are interested in clustering observations so that we have two folding methods, $I\times(T\times P)$ and $I\times(P\times T)$ to investigate.

A kernel function for the kernel principal component analysis is required to be at least semi-positive definite such as polynomial kernel function (PKF), radial basis function (RBF), normal kernel function, and so on (Scholkopf and Smola 2002). The general question of how to choose the ideal kernel for a given data set is an open problem and is not discussed here. In this study, polynomial and radial basis functions had been tried, and the radial basis function was selected since RBF was found to work better in clustering our complex data. The parameter for RBF, c, will be investigated since the parameter seems to have relations with a given data structure.

We used the *k*-means algorithm, where each cluster is represented by the mean value of the objects in the cluster, as a partitioning method. This algorithm works better for larger data sets as this simulation study was designed than conventional hierarchical clustering methods (Johnson and Wichern 2002). We specified two clusters in advance. We do not attempt to figure out which clustering method might be promising in this study since the scores we obtained from KMPCA are nothing but static data. And, these are well explained in every text.

4.1 Generating non-linear multivariate time series mean profiles

An $i^{th}$ sample observation of a univariate time series dataset with $T$ temporal points and $P$ variables can be conveniently represented by the following matrix expression:

$$X_i = \begin{bmatrix} x_{i11} & x_{i12} & \cdots & x_{i1T} \\ x_{i21} & x_{i22} & \cdots & x_{i2T} \\ \vdots & \vdots & \ddots & \vdots \\ x_{iP1} & x_{iP2} & \cdots & x_{iPT} \end{bmatrix}, \ , i = 1, 2, \ldots, N \qquad (4.1)$$

where $x_{ipt}$ is the measurement of the $i^{th}$ object's $p^{th}$ variable at the $t^{th}$ time point.

As defined before, two time series are similar if they have enough non-overlapping time-ordered subsequences that are similar. That is, if there is a group of time series in the same cluster, the time series has a similar pattern: they have close enough values for each time point. If we overlap the time series in the same plot, they could be expressed as a time series with the same mean profile and error: $X = M$ (Model structure) + $E$ (Error structure). In other words, if two time series have different mean profiles, then these are two different time series.

However, two different time series can have different error structure also. Suppose two time series have same mean profiles but one has pure error for each time points so that the data are close to the mean profiles and the other has much bigger error or extraneous effects which pull or push the data from or to the mean profiles. These time series should be regarded as two

different groups of time series. The factors which affect the error structure will be auto-correlation, cross-correlation, or both when we handle multivariate time series.

### 4.1.1 Model structure

We generated two different nonlinear tri-variate time series data sets with various combinations of auto-correlations and cross-correlations. Using the concept $X = M$ (Model structure) + $E$ (Error structure), we first built two different $M$ and $L$ matrices:

$$M = \begin{bmatrix} \mu_{11} & \mu_{21} & \cdots & \mu_{P1} \\ \mu_{12} & \mu_{22} & \cdots & \mu_{P2} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{1T} & \mu_{2T} & \cdots & \mu_{PT} \end{bmatrix} \text{ and } L = \begin{bmatrix} \tau_{11} & \tau_{21} & \cdots & \tau_{P1} \\ \tau_{12} & \tau_{22} & \cdots & \tau_{P2} \\ \vdots & \vdots & \ddots & \vdots \\ \tau_{1T} & \tau_{2T} & \cdots & \tau_{PT} \end{bmatrix} \tag{4.2}$$

where $\mu_{kt}$ is the mean of the $p^{\text{th}}$ variable profile at time period $t$, for $p$ = 1, 2, 3, ..., $P$ and $t$ = 1, 2, ..., $T$. So, each column vector is a mean profile of a time series.

For the mean profiles, we used the three-variable batch process data from a statistical process control study (Kim and Adams 2009). Using the time series plots in the paper, we generated thirty mean values for each time series with a nonlinear function which is used by Warner and Misra (1996) as follows:

$$\mu_{pt} = b_{1p} \exp^{b_{2p}\frac{t}{T}}[\ln(b_{3p}\frac{t}{T} + b_{4p}) + b_{5p}] + b_{6p}, \text{ where } t = 1, 2, ..., T. \tag{4.3}$$

The next table displays the parameters of two different mean profiles with three variable time series each with 30 time points. For the cases for shorter temporal point, the same mean profiles were used from the first time point to the number the case needs. We kept front part from the original mean profiles since the values converge on a certain value as number of temporal points increase.

Table 4.1. Parameters of Six Different Models for Simulation Studies

| Parameter | $b_1$ | $b_2$ | $b_3$ | $b_4$ | $b_5$ | $b_6$ |
|---|---|---|---|---|---|---|
| Model 1 | 2.50 | -10.00 | 10.00 | -1.00 | 1.00 | 0.25 |
| Model 2 | -5.00 | -8.00 | 1.00 | 0.20 | 1.29 | 0.50 |
| Model 3 | 2.00 | -10.00 | 1.00 | 0.20 | 1.50 | -0.10 |
| Model 4 | 2.00 | -10.00 | 1.00 | 0.20 | -1.00 | 1.00 |
| Model 5 | 2.00 | -10.00 | 1.10 | 0.20 | -1.00 | 2.00 |
| Model 6 | 2.00 | -10.00 | 0.90 | 0.20 | -1.00 | 0.00 |

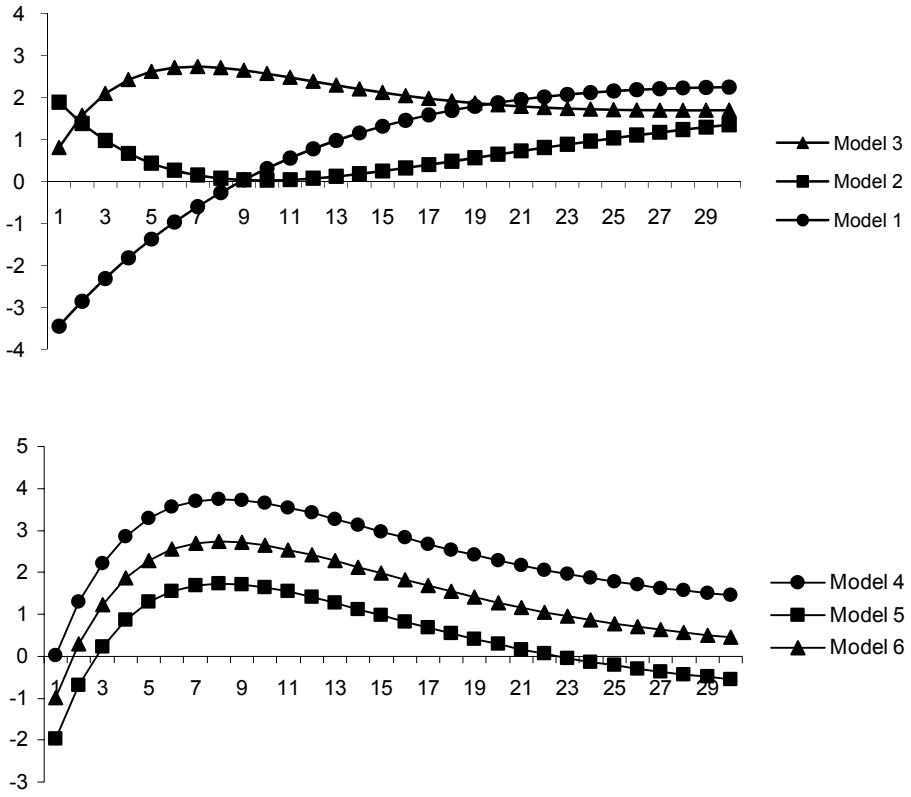


Figure 4.2. Two Different Mean Profiles for Generating Simulation Data

Model 1 through 3 will comprise the group 1, and 4 through 6 will do group 2. For the simulation study I, we used group 1 as tri-variate mean profiles, and both group 1 and 2 were used for the simulation study II to compare two different mean groups with same error structure.

4.1.2 Error structure

Let $x_{pt}$ be an observation of the $p^{th}$ variable at time period $t$ and define $e_{pt}$ to be $x_{pt} - \mu_{pt}$. Then, the matrix $E$ is defined by

$$E = \begin{bmatrix} e_{11} & e_{21} & \cdots & e_{P1} \\ e_{12} & e_{22} & \cdots & e_{P2} \\ \vdots & \vdots & \ddots & \vdots \\ e_{1T} & e_{2T} & \cdots & e_{PT} \end{bmatrix}, \text{ where } \underline{E}_p \text{ is } 1 \times T \text{, the } p^{\text{th}} \text{ vector.} \quad (4.4)$$

Depending on the variance structure of matrix $E$, a time series is considered different series from others. In this study, two different types of correlations are considered for generating error matrices: cross-correlation and auto-correlation at time period $t$.

4.1.3 Auto-correlation and cross correlation

To generate a time series with cross correlation only, $X$ is defined as $X = M + E$, where $E_t \sim MN(0, \Sigma_E)$ and $\Sigma_E$ are given by

$\Sigma_E = \alpha \cdot diag\Sigma_M + \Sigma_{\text{cov}}$, where $0 < \alpha < 1$,

$$\Sigma_{COV} = \begin{bmatrix} 0 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1P} \\ \sigma_{21} & 0 & \sigma_{23} & \cdots & \sigma_{2P} \\ \sigma_{31} & \sigma_{32} & 0 & \cdots & \sigma_{3P} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{P1} & \sigma_{P2} & \sigma_{P3} & \cdots & \sigma_{PP} \end{bmatrix}, \quad (4.5)$$

and $\Sigma_M$ is diagonal matrix and the variance-covariance matrix of $M$.

The $\alpha$ size depends on how stable the model which a time series might have is. If one could control the model to be stable, then $\alpha$ should be small enough such as less than 0.1. If the $\alpha$ size is close to 0.5, the model, so-called "white-noise" or "random walks," is not valid any more. In real, most cases of data are somewhat noisy. Therefore, we attempted to see how the proposed method works when the data has different level of error variance from $\alpha$ = 0.05 to 0.4.

The proportion of error variance to total variance is depending on auto-correlation, and therefore will be considered with auto-correlation coefficients simultaneously.

For the example of calculation when $\alpha = 0.1$, $diag\Sigma_E = 0.1 \cdot diag\Sigma_M$. Since

$$\rho_{X1,X2} = \frac{Cov(x1, x2)}{\sigma_{x1}\sigma_{x2}}, \tag{4.6}$$

if an arbitrary correlation matrix, $\Re$, is set, off-diagonal elements of $\Sigma_{cov}$ are obtained. In this study, three arbitrary sets of $\Re$ with strong, moderate, and weak coefficients of cross correlation between the variables are considered. The corresponding $\Sigma_{cov}$ is given by the following:

$$\Re_1 = \begin{bmatrix} 1 & 0.8 & 0.7 \\ 0.8 & 1 & 0.8 \\ 0.7 & 0.8 & 1 \end{bmatrix}. \tag{4.7}$$

Before calculation, determinants are checked to keep invertibility of the matrices.

For generating a time series with large coefficients of cross correlation,

$$\alpha \cdot diag\Sigma_M = 0.1 \times \begin{bmatrix} Var(X_1) & 0 & 0 \\ 0 & Var(X_2) & 0 \\ 0 & 0 & Var(X_3) \end{bmatrix}.$$

$$\Sigma_{COV} = \begin{bmatrix} 0 & 0.8 \times \sqrt{Var(X_1)} \times \sqrt{Var(X_2)} & 0.7 \times \sqrt{Var(X_1)} \times \sqrt{Var(X_2)} \\ 0.8 \times \sqrt{Var(X_2)} \times \sqrt{Var(X_1)} & 0 & 0.8 \times \sqrt{Var(X_2)} \times \sqrt{Var(X_3)} \\ 0.7 \times \sqrt{Var(X_3)} \times \sqrt{Var(X_1)} & 0.8 \times \sqrt{Var(X_3)} \times \sqrt{Var(X_2)} & 0 \end{bmatrix} \tag{4.8}$$

Hence, $\Sigma_{E11} = \alpha \cdot diag\Sigma_{M1} + \Sigma_{cov11} =$

$$\begin{bmatrix} 0.1 \times Var(X_1) & 0.8 \times \sqrt{Var(X_1)} \times \sqrt{Var(X_2)} & 0.7 \times \sqrt{Var(X_1)} \times \sqrt{Var(X_2)} \\ 0.8 \times \sqrt{Var(X_2)} \times \sqrt{Var(X_1)} & 0.1 \times Var(X_2) & 0.8 \times \sqrt{Var(X_2)} \times \sqrt{Var(X_3)} \\ 0.7 \times \sqrt{Var(X_3)} \times \sqrt{Var(X_1)} & 0.8 \times \sqrt{Var(X_3)} \times \sqrt{Var(X_2)} & 0.1 \times Var(X_3) \end{bmatrix}. \tag{4.9}$$

For generating time series data with moderate cross correlation coefficients, the following cross correlation structure is used. In here, the extent of cross correlation is set as moderate relative to first case.

$$\Re_2 = \begin{bmatrix} 1 & 0.4 & 0.5 \\ 0.4 & 1 & 0.4 \\ 0.5 & 0.4 & 1 \end{bmatrix} \tag{4.10}$$

For generating time series data with weak cross correlation coefficients, the following cross correlation structure is used. Also, this cross-correlation is relatively lower than former two cases so that it is set as low cross-correlation.

$$\Re_3 = \begin{bmatrix} 1 & 0.1 & 0.1 \\ 0.1 & 1 & -0.1 \\ 0.1 & -0.1 & 1 \end{bmatrix} \tag{4.11}$$

To give auto-correlation to the generating data, we used the Box-Jenkins Auto-Regressive and Moving Average with first order (ARMA (1, 1)). Instead of handling ARIMA time series directly, a preprocessing step of differencing was first applied to convert each nonstationary ARIMA time series into the corresponding stationary time series (Shumway 1988). Therefore, ARMA with first order model is used in this study.

By adjusting parameters of the function below, different error structures are generated as follows:

$$ARMA\ (1,\ 1) = X_t = \begin{bmatrix} \phi_{1,1} \\ \phi_{1,2} \\ \vdots \\ \phi_{1,p} \end{bmatrix} \begin{bmatrix} x_{1,t-1} & x_{2,t-2} & \cdots & x_{p,t-1} \end{bmatrix} + \begin{bmatrix} 1 & \theta_{1,1} \\ 1 & \theta_{1,2} \\ \vdots & \vdots \\ 1 & \theta_{1,p} \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t} & \varepsilon_{2,t} & \cdots & \varepsilon_{p,t} \\ \varepsilon_{1,t-1} & \varepsilon_{2,t-1} & \cdots & \varepsilon_{p,t-1} \end{bmatrix} \tag{4.12}$$

And, the error matrix $E_t$ is $E_t \sim IID(0, \Sigma)$,

$$\text{where } \Sigma = \begin{bmatrix} Var(X_{1,t}) & Cov(X_{1,t}, X_{2,t}) & Cov(X_{1,t}, X_{3,t}) & \cdots & Cov(X_{1,t}, X_{p,t}) \\ Cov(X_{2,t}, X_{1,t}) & Var(X_{2,t}) & Cov(X_{2,t}, X_{3,t}) & \cdots & Cov(X_{2,t}, X_{p,t}) \\ Cov(X_{3,t}, X_{1,t}) & Cov(X_{3,t}, X_{2,t}) & Var(X_{3,t}) & \cdots & Cov(X_{3,t}, X_{p,t}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Cov(X_{p,t}, X_{1,t}) & Cov(X_{p,t}, X_{2,t}) & Cov(X_{p,t}, X_{3,t}) & \cdots & Var(X_{p,t}) \end{bmatrix} \quad (4.13)$$

For the stationary assumption, we checked the constraints of the model. Let us consider a univariate time series of ARMA (1, 1) temporarily for convenience.

$y_t = \theta_0 + \phi_1 y_{t-1} + \varepsilon_t - \theta_1 \varepsilon_{t-1}$ where $\theta_0$ is an initial value for the model.

For the mean, $E(y_t) = \theta_0 + \phi_1 E(y_{t-1})$, therefore, $\mu_t = \theta_0 + \phi_1 \mu_t$ and $(1-\phi_1)\mu_t = \theta_0$. So,

the total mean is $\mu = \dfrac{\theta_0}{1-\phi_1}$ . (4.14)

Thus, if we set up the initial values for the generated model all the same as zero, the mean value for the series is zero, which is used in this simulation study.

For the variance, $Var(y_t) = \phi_1^2 Var(y_{t-1}) + Var(\varepsilon_t) + \theta_1^2 Var(\varepsilon_{t-1})$ where $Var(x_t) = Var(x_{t-1}) = \sigma_x$ since it is a stationary time series. According to the error variance assumptions, $Var(\varepsilon_t) = \sigma_{error}$ is error variance at point *t* and $Var(\theta_t) = Var(\theta_{t-1}) = \sigma_{error}$. Also, errors for time points are independent of each other and orthogonal to the model structure variables.

So, $(1-\phi_1^2)\sigma_{Total}{}^2 = (1+\theta_1^2)\sigma_{Error}^2$. Therefore, the result is $\sigma_{Total}^2 = \dfrac{1+\theta_1^2}{1-\phi_1^2}\sigma_{Error}^2$. (4.15)

The error variance is proportion of the total variance of a model, and the proportion is composed of the coefficients of the orders of AR and MA components. That is, total variance and error variance depend on the coefficients of the ARMA model. To set up the auto-correlation

coefficients for various simulation data, we need to seek the area of combinations calculating the proportion. For example, for $\alpha$ = 0.1, error variance is 10 percent of total variance;

$$\frac{1+\theta_1^2}{1-\phi_1^2} = 0.1. \tag{4.16}$$

With stationary assumption, $|\phi_1| \le 1$, and invertibility assumption, $|\theta_1| \le 1$, there are combinations of proportion of error variance, AR coefficients, and MA coefficients.

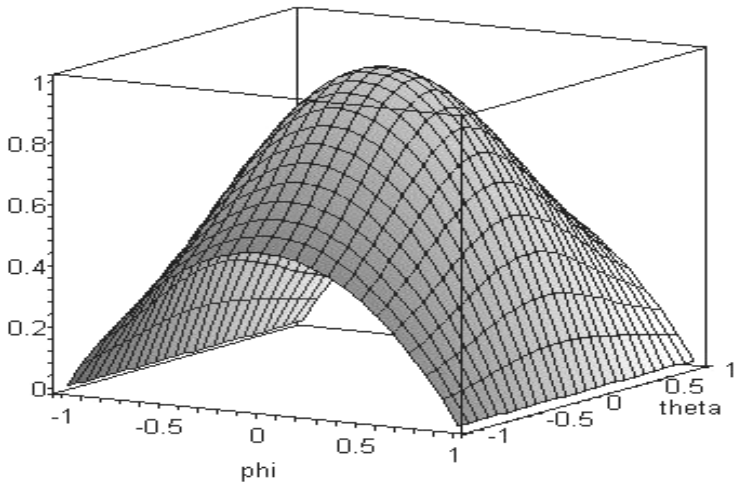


Figure 4.3. 3-D Graph of $\alpha$, $\phi_1$, and $\theta_1$

So, we sought possible area of simulation from the 3-D graph of these coefficients as in Figure 4.3. AR coefficient changes much according to the proportion of error variance while MA coefficient does not. Therefore, we try to fix $\theta_1$ with a number close to as big as 0.9 and change $\phi_1$ so that we can target the $\alpha$ size as small as possible. One thing notable here that we don't consider weak auto-correlation case since it causes a time series as noisy as the series might be white noise which we are not interested in clustering. We chose several cases from the possible area.

We built two tri-variate mean profile model structures using a nonlinear function with six different parameters. For the error structure, we considered various auto-correlation and cross-correlation coefficients corresponding to the error variance proportion. Using the variance obtained from generated data, we calculated error variance covariance matrices. After that, we weighted the error variance covariance matrices with different error variance proportion. Finally, the next table displays several cases to look through.

Table 4.2. Model Cases for Simulation with Various Combinations of AR Coefficients, MA Coefficients, Error Variance Proportion, and Cross Correlation Structures

| Case Number | Model Stability | AR Coefficients | MA Coefficients | Cross correlation matrix |
|---|---|---|---|---|
| 1 | $\alpha = 0.1$ | $\phi_1 = 0.904986$ | $\theta_1 = 0.9$ | $\Re_1 = \begin{bmatrix} 1 & .1 & .1 \\ .1 & 1 & -.1 \\ .1 & -.1 & 1 \end{bmatrix}$ |
| 2 | $\alpha = 0.2$ | $\phi_1 = 0.798748$ | $\theta_1 = 0.9$ | $\Re_1 = \begin{bmatrix} 1 & .1 & .1 \\ .1 & 1 & -.1 \\ .1 & -.1 & 1 \end{bmatrix}$ |
| 3 | $\alpha = 0.3$ | $\phi_1 = 0.676018$ | $\theta_1 = 0.9$ | $\Re_1 = \begin{bmatrix} 1 & .1 & .1 \\ .1 & 1 & -.1 \\ .1 & -.1 & 1 \end{bmatrix}$ |
| 4 | $\alpha = 0.4$ | $\phi_1 = 0.525356$ | $\theta_1 = 0.9$ | $\Re_1 = \begin{bmatrix} 1 & .1 & .1 \\ .1 & 1 & -.1 \\ .1 & -.1 & 1 \end{bmatrix}$ |
| 5 | $\alpha = 0.1$ | $\phi_1 = 0.904986$ | $\theta_1 = 0.9$ | $\Re_2 = \begin{bmatrix} 1 & .4 & .5 \\ .4 & 1 & .4 \\ .5 & .4 & 1 \end{bmatrix}$ |
| 6 | $\alpha = 0.2$ | $\phi_1 = 0.798748$ | $\theta_1 = 0.9$ | $\Re_2 = \begin{bmatrix} 1 & .4 & .5 \\ .4 & 1 & .4 \\ .5 & .4 & 1 \end{bmatrix}$ |
| 7 | $\alpha = 0.3$ | $\phi_1 = 0.676018$ | $\theta_1 = 0.9$ | $\Re_2 = \begin{bmatrix} 1 & .4 & .5 \\ .4 & 1 & .4 \\ .5 & .4 & 1 \end{bmatrix}$ |

| | | | | |
|---|---|---|---|---|
| 8 | $\alpha = 0.4$ | $\phi_1 = 0.525356$ | $\theta_1 = 0.9$ | $\Re_2 = \begin{bmatrix} 1 & .4 & .5 \\ .4 & 1 & .4 \\ .5 & .4 & 1 \end{bmatrix}$ |
| 9 | $\alpha = 0.1$ | $\phi_1 = 0.904986$ | $\theta_1 = 0.9$ | $\Re_3 = \begin{bmatrix} 1 & .7 & .8 \\ .7 & 1 & .7 \\ .8 & .7 & 1 \end{bmatrix}$ |
| 10 | $\alpha = 0.2$ | $\phi_1 = 0.798748$ | $\theta_1 = 0.9$ | $\Re_3 = \begin{bmatrix} 1 & .7 & .8 \\ .7 & 1 & .7 \\ .8 & .7 & 1 \end{bmatrix}$ |
| 11 | $\alpha = 0.3$ | $\phi_1 = 0.676018$ | $\theta_1 = 0.9$ | $\Re_3 = \begin{bmatrix} 1 & .7 & .8 \\ .7 & 1 & .7 \\ .8 & .7 & 1 \end{bmatrix}$ |
| 12 | $\alpha = 0.4$ | $\phi_1 = 0.525356$ | $\theta_1 = 0.9$ | $\Re_3 = \begin{bmatrix} 1 & .7 & .8 \\ .7 & 1 & .7 \\ .8 & .7 & 1 \end{bmatrix}$ |

#### 4.1.4 Temporal points

Time series data can be distinguished by the number of temporal points since the number can affect the error structure as well as the mean profiles. So, we investigated the clustering performances with different numbers of temporal points.

For the valid and stable time series modeling, temporal points should be large enough such as over 30. However, in a real world, short-time period time series prevail especially in clinical, medical, or health care management fields. And, we are trying to focus on the data of these fields in this study eventually. So, the number of temporal points was limited to less than 30 points. In the clinical trial field, the least number of repeated measures to obtain validity of test is six. So, we started with six temporal points for the simulation data. We arbitrary selected the middle point, 15 to avoid any quadratic effects of this factor. Therefore, we generated 6, 15, and 30 temporal points of data for the simulation.

The simulation study is outlined by the following comparisons:

1) Comparing one mean trajectory group with pure error only as the reference group to the groups with same mean trajectory with nine different combinations of cross-correlation and auto-correlation.
2) Comparing two different mean trajectory groups with same cross-correlation and auto-correlation. Nine different combinations of cross-correlation and auto-correlation will be simulated and tested.

## 4.2 Simulation study I

The first simulation study is to investigate if the proposed method would tell multivariate time series with pure error (error variance proportion, $\alpha = 0.1$) from those with the error structure of cross-correlation and auto-correlation. These simulation results will guide practitioners to apply KMPCA to the multivariate time series data to cluster as well as enlighten the classification researchers to know this proposed method more when it is used for clustering multivariate time series data.

This would be close to real world cases; for example, most credit applicants in the same group such as good credit, bad credit, or default group would have similar financial records for the time being. If an credit customer has had financial problems at some point, his record would start to show dropped payment history or irregular payment time, and so on, which might be a trigger to transit from bad credit group to default group.

### 4.2.1 Reference group

The reference group has the first mean profiles, model 1, 2, and 3 as in the figure 4. 2. The group is consistently compared to the control group which has the same mean with different combinations of error structure. So, we can conduct a hypothesis test whether a multivariate time

series observation is included in the reference group or not as well as compare the average clustering success rate of a control group case to that of another. We gave 10 percents margin of error for each temporal point to confirm not having extreme values.

### 4.2.2 Control groups

We generated control groups to compare the performances of the proposed method. These have the same mean profile but different error variance scheme. Three different cross-correlations (weak, moderate, and strong) were multiplied to the estimated variance-covariance matrix. After that, three different auto-correlations ($\phi_1 = 0.904986$, 0.798748, and 0.676018) were also multiplied to the matrices. We dropped the least case, $\phi_1 = 0.525356$, which showed not much difference with the last case from the preliminary simulation tests. Thus, nine different cases were generated for three cases of temporal points.

For the parameters of the kernel function, c = 2 showed good results of similar studies (Liu, et al. 2005). We arbitrary selected c = 1 to compare the performances between these two parameters to check if there is any effect on the clustering performance. In addition, two folding methods for the MPCA were considered. Therefore, a total number of cases was 108 (3×3×3×2×2) for the control groups as in table 4. 3.

Table 4.3. All the Cases of Factors and Levels Considered in the Simulation Studies

| | Factors and levels | | | | |
|---|---|---|---|---|---|
| Case Number | Folding Method | RBF parameter, c | Number of temporal points | Cross-correlation | Auto-correlation |
| 1 | I×(J×K) | c =2 | t = 6 | Weak | $\phi_1 = 0.68$ |
| 2 | I×(J×K) | c =2 | t = 6 | Weak | $\phi_1 = 0.79$ |
| 3 | I×(J×K) | c =2 | t = 6 | Weak | $\phi_1 = 0.9$ |
| 4 | I×(J×K) | c =2 | t = 6 | Moderate | $\phi_1 = 0.68$ |

| | | | | | |
|---|---|---|---|---|---|
| 5 | I×(J×K) | c =2 | t = 6 | Moderate | $\phi_1 = 0.79$ |
| 6 | I×(J×K) | c =2 | t = 6 | Moderate | $\phi_1 = 0.9$ |
| 7 | I×(J×K) | c =2 | t = 6 | Strong | $\phi_1 = 0.68$ |
| 8 | I×(J×K) | c =2 | t = 6 | Strong | $\phi_1 = 0.79$ |
| 9 | I×(J×K) | c =2 | t = 6 | Strong | $\phi_1 = 0.9$ |
| 10 | I×(J×K) | c =2 | t = 15 | Weak | $\phi_1 = 0.68$ |
| 11 | I×(J×K) | c =2 | t = 15 | Weak | $\phi_1 = 0.79$ |
| 12 | I×(J×K) | c =2 | t = 15 | Weak | $\phi_1 = 0.9$ |
| 13 | I×(J×K) | c =2 | t = 15 | Moderate | $\phi_1 = 0.68$ |
| 14 | I×(J×K) | c =2 | t = 15 | Moderate | $\phi_1 = 0.79$ |
| 15 | I×(J×K) | c =2 | t = 15 | Moderate | $\phi_1 = 0.9$ |
| 16 | I×(J×K) | c =2 | t = 15 | Strong | $\phi_1 = 0.68$ |
| 17 | I×(J×K) | c =2 | t = 15 | Strong | $\phi_1 = 0.79$ |
| 18 | I×(J×K) | c =2 | t = 15 | Strong | $\phi_1 = 0.9$ |
| 19 | I×(J×K) | c =2 | t = 30 | Weak | $\phi_1 = 0.68$ |
| 20 | I×(J×K) | c =2 | t = 30 | Weak | $\phi_1 = 0.79$ |
| 21 | I×(J×K) | c =2 | t = 30 | Weak | $\phi_1 = 0.9$ |
| 22 | I×(J×K) | c =2 | t = 30 | Moderate | $\phi_1 = 0.68$ |
| 23 | I×(J×K) | c =2 | t = 30 | Moderate | $\phi_1 = 0.79$ |
| 24 | I×(J×K) | c =2 | t = 30 | Moderate | $\phi_1 = 0.9$ |
| 25 | I×(J×K) | c =2 | t = 30 | Strong | $\phi_1 = 0.68$ |
| 26 | I×(J×K) | c =2 | t = 30 | Strong | $\phi_1 = 0.79$ |
| 27 | I×(J×K) | c =2 | t = 30 | Strong | $\phi_1 = 0.9$ |
| 28 | I×(J×K) | c =1 | t = 6 | Weak | $\phi_1 = 0.68$ |
| 29 | I×(J×K) | c =1 | t = 6 | Weak | $\phi_1 = 0.79$ |
| 30 | I×(J×K) | c =1 | t = 6 | Weak | $\phi_1 = 0.9$ |

| | | | | | |
|---|---|---|---|---|---|
| 31 | I×(J×K) | c =1 | t = 6 | Moderate | $\phi_1 = 0.68$ |
| 32 | I×(J×K) | c =1 | t = 6 | Moderate | $\phi_1 = 0.79$ |
| 33 | I×(J×K) | c =1 | t = 6 | Moderate | $\phi_1 = 0.9$ |
| 34 | I×(J×K) | c =1 | t = 6 | Strong | $\phi_1 = 0.68$ |
| 35 | I×(J×K) | c =1 | t = 6 | Strong | $\phi_1 = 0.79$ |
| 36 | I×(J×K) | c =1 | t = 6 | Strong | $\phi_1 = 0.9$ |
| 37 | I×(J×K) | c =1 | t = 15 | Weak | $\phi_1 = 0.68$ |
| 38 | I×(J×K) | c =1 | t = 15 | Weak | $\phi_1 = 0.79$ |
| 39 | I×(J×K) | c =1 | t = 15 | Weak | $\phi_1 = 0.9$ |
| 40 | I×(J×K) | c =1 | t = 15 | Moderate | $\phi_1 = 0.68$ |
| 41 | I×(J×K) | c =1 | t = 15 | Moderate | $\phi_1 = 0.79$ |
| 42 | I×(J×K) | c =1 | t = 15 | Moderate | $\phi_1 = 0.9$ |
| 43 | I×(J×K) | c =1 | t = 15 | Strong | $\phi_1 = 0.68$ |
| 44 | I×(J×K) | c =1 | t = 15 | Strong | $\phi_1 = 0.79$ |
| 45 | I×(J×K) | c =1 | t = 15 | Strong | $\phi_1 = 0.9$ |
| 46 | I×(J×K) | c =1 | t = 30 | Weak | $\phi_1 = 0.68$ |
| 47 | I×(J×K) | c =1 | t = 30 | Weak | $\phi_1 = 0.79$ |
| 48 | I×(J×K) | c =1 | t = 30 | Weak | $\phi_1 = 0.9$ |
| 49 | I×(J×K) | c =1 | t = 30 | Moderate | $\phi_1 = 0.68$ |
| 50 | I×(J×K) | c =1 | t = 30 | Moderate | $\phi_1 = 0.79$ |
| 51 | I×(J×K) | c =1 | t = 30 | Moderate | $\phi_1 = 0.9$ |
| 52 | I×(J×K) | c =1 | t = 30 | Strong | $\phi_1 = 0.68$ |
| 53 | I×(J×K) | c =1 | t = 30 | Strong | $\phi_1 = 0.79$ |
| 54 | I×(J×K) | c =1 | t = 30 | Strong | $\phi_1 = 0.9$ |
| 55 | I×(K×J) | c =2 | t = 6 | Weak | $\phi_1 = 0.68$ |
| 56 | I×(K×J) | c =2 | t = 6 | Weak | $\phi_1 = 0.79$ |

| | | | | | |
|---|---|---|---|---|---|
| 57 | I×(K×J) | c =2 | t = 6 | Weak | $\phi_1 = 0.9$ |
| 58 | I×(K×J) | c =2 | t = 6 | Moderate | $\phi_1 = 0.68$ |
| 59 | I×(K×J) | c =2 | t = 6 | Moderate | $\phi_1 = 0.79$ |
| 60 | I×(K×J) | c =2 | t = 6 | Moderate | $\phi_1 = 0.9$ |
| 61 | I×(K×J) | c =2 | t = 6 | Strong | $\phi_1 = 0.68$ |
| 62 | I×(K×J) | c =2 | t = 6 | Strong | $\phi_1 = 0.79$ |
| 63 | I×(K×J) | c =2 | t = 6 | Strong | $\phi_1 = 0.9$ |
| 64 | I×(K×J) | c =2 | t = 15 | Weak | $\phi_1 = 0.68$ |
| 65 | I×(K×J) | c =2 | t = 15 | Weak | $\phi_1 = 0.79$ |
| 66 | I×(K×J) | c =2 | t = 15 | Weak | $\phi_1 = 0.9$ |
| 67 | I×(K×J) | c =2 | t = 15 | Moderate | $\phi_1 = 0.68$ |
| 68 | I×(K×J) | c =2 | t = 15 | Moderate | $\phi_1 = 0.79$ |
| 69 | I×(J×K) | c =2 | t = 15 | Moderate | $\phi_1 = 0.9$ |
| 70 | I×(K×J) | c =2 | t = 15 | Strong | $\phi_1 = 0.68$ |
| 71 | I×(K×J) | c =2 | t = 15 | Strong | $\phi_1 = 0.79$ |
| 72 | I×(K×J) | c =2 | t = 15 | Strong | $\phi_1 = 0.9$ |
| 73 | I×(K×J) | c =2 | t = 30 | Weak | $\phi_1 = 0.68$ |
| 74 | I×(K×J) | c =2 | t = 30 | Weak | $\phi_1 = 0.79$ |
| 75 | I×(K×J) | c =2 | t = 30 | Weak | $\phi_1 = 0.9$ |
| 76 | I×(K×J) | c =2 | t = 30 | Moderate | $\phi_1 = 0.68$ |
| 77 | I×(K×J) | c =2 | t = 30 | Moderate | $\phi_1 = 0.79$ |
| 78 | I×(K×J) | c =2 | t = 30 | Moderate | $\phi_1 = 0.9$ |
| 79 | I×(K×J) | c =2 | t = 30 | Strong | $\phi_1 = 0.68$ |
| 80 | I×(K×J) | c =2 | t = 30 | Strong | $\phi_1 = 0.79$ |
| 81 | I×(K×J) | c =2 | t = 30 | Strong | $\phi_1 = 0.9$ |
| 82 | I×(K×J) | c =1 | t = 6 | Weak | $\phi_1 = 0.68$ |

| | | | | | |
|---|---|---|---|---|---|
| 83 | I×(K×J) | c =1 | t = 6 | Weak | $\phi_1 = 0.79$ |
| 84 | I×(K×J) | c =1 | t = 6 | Weak | $\phi_1 = 0.9$ |
| 85 | I×(K×J) | c =1 | t = 6 | Moderate | $\phi_1 = 0.68$ |
| 86 | I×(J×K) | c =1 | t = 6 | Moderate | $\phi_1 = 0.79$ |
| 87 | I×(K×J) | c =1 | t = 6 | Moderate | $\phi_1 = 0.9$ |
| 88 | I×(K×J) | c =1 | t = 6 | Strong | $\phi_1 = 0.68$ |
| 89 | I×(K×J) | c =1 | t = 6 | Strong | $\phi_1 = 0.79$ |
| 90 | I×(K×J) | c =1 | t = 6 | Strong | $\phi_1 = 0.9$ |
| 91 | I×(K×J) | c =1 | t = 15 | Weak | $\phi_1 = 0.68$ |
| 92 | I×(K×J) | c =1 | t = 15 | Weak | $\phi_1 = 0.79$ |
| 93 | I×(K×J) | c =1 | t = 15 | Weak | $\phi_1 = 0.9$ |
| 94 | I×(K×J) | c =1 | t = 15 | Moderate | $\phi_1 = 0.68$ |
| 95 | I×(J×K) | c =1 | t = 15 | Moderate | $\phi_1 = 0.79$ |
| 96 | I×(K×J) | c =1 | t = 15 | Moderate | $\phi_1 = 0.9$ |
| 97 | I×(K×J) | c =1 | t = 15 | Strong | $\phi_1 = 0.68$ |
| 98 | I×(K×J) | c =1 | t = 15 | Strong | $\phi_1 = 0.79$ |
| 99 | I×(K×J) | c =1 | t = 15 | Strong | $\phi_1 = 0.9$ |
| 100 | I×(K×J) | c =1 | t = 30 | Weak | $\phi_1 = 0.68$ |
| 101 | I×(K×J) | c =1 | t = 30 | Weak | $\phi_1 = 0.79$ |
| 102 | I×(K×J) | c =1 | t = 30 | Weak | $\phi_1 = 0.9$ |
| 103 | I×(K×J) | c =1 | t = 30 | Moderate | $\phi_1 = 0.68$ |
| 104 | I×(J×K) | c =1 | t = 30 | Moderate | $\phi_1 = 0.79$ |
| 105 | I×(K×J) | c =1 | t = 30 | Moderate | $\phi_1 = 0.9$ |
| 106 | I×(K×J) | c =1 | t = 30 | Strong | $\phi_1 = 0.68$ |
| 107 | I×(K×J) | c =1 | t = 30 | Strong | $\phi_1 = 0.79$ |
| 108 | I×(K×J) | c =1 | t = 30 | Strong | $\phi_1 = 0.9$ |

4.2.3 Determining size of simulation

In this simulation study, we have only two groups, so whether a given observation would be from the reference group or not is a binomial distribution ~ b(NSIM, $\pi$) where $\pi$ = P[A certain observation belongs to the reference group] and NSIM is the number of simulation. However, we do not know true proportion of success, which is how many observations are included in the targeted group in the population. So, to determine the size of simulation needs a restriction.

The margin of error in estimating the proportion of success in clustering control group from the reference group is determined in advance as 0.05. Therefore, the simulation size, NSIM, is obtained to meet the specified margin of error by

$$NSIM = \hat{P} \cdot (1 - \hat{P}) \cdot \left( \frac{z_{\alpha/2}}{0.05} \right)^2 . \tag{4.17}$$

Since $\hat{P}$ is unknown, the most conservative approach is used for determining NSIM by given value of $\hat{P}$=0.5 with $\alpha = 0.05$.

$$\text{Then, } NSIM = \hat{P} \cdot (1 - \hat{P}) \cdot \left( \frac{z_{\alpha/2}}{0.05} \right)^2 = 0.25 \left( \frac{1.96}{0.05} \right)^2 \approx 385 . \tag{4.18}$$

Hence, the sample size obtained is generally larger than necessary and the margin of error less than required.

4.2.4 Number of the observations for the groups ($n_1 = n_2$)

Now we have two groups, reference group, R, and control group, C. We assume that $\alpha_R$ = Probability that an observation of R would be assigned to C and $\alpha_C$ = Probability that an observation of C would be assigned to R. To obtain the required sample numbers of both groups, we need to fix type I error sizes for both groups. If we increase the number of observation for

each group, these type I errors can converge on zero according to central limit theorem. Then the clustering success rates can be converged on the true proportion of the population.

However, increasing sample number causes tremendous calculating time. And, kernel methods have limitations in calculating large number of observations. Therefore, we need to compromise appropriate numbers enough to fulfill this simulation study.

We set up targeted numbers as 0.05 for both errors. Then total error is around 0.1 (1 – (.95× .95 = .9025) = 0.0975). Now, we selected a simulation case among 108 cases to seek number of observations. The temporal point is set as six for shorter computing time, and the folding method is set as I×(K×T). As the number of observations were increased as $n_1 = n_2 = 100$, 200, 400, and 600, the success rates of clustering for nine different cases were checked.

The table 4.4 revealed that the clustering success rates were not so much different as the number of observations for both groups increases. The differences of success rates were tested by paired t-tests between the number of observations of two groups is 100 and that of 200, between the groups of 200 and those of 400, and between the groups of 400 and those of 600. We found positive increases in most cases when the observation increases from 100 to 200 even though there are no statistically significant test results. Therefore, we chose 200 for each group for the perspective of the efficiency of calculating time for the simulation study.

Table 4.4. Clustering Success Rates as Number of Observations for Both Reference and Control Group Increase

| Case compared | n1 = n2 = 100 | n1 = n2 = 200 | n1 = n2 = 400 | n1 = n2 = 600 |
|---|---|---|---|---|
| Case 1 | 92.800 | 93.080 | 93.237 | 93.336 |
| Case 2 | 80.610 | 80.785 | 81.137 | 81.961 |
| Case 3 | 50.750 | 50.518 | 50.420 | 50.280 |

| | | | | |
|---|---|---|---|---|
| Case 4 | 90.670 | 90.935 | 91.219 | 94.368 |
| Case 5 | 78.040 | 78.115 | 78.284 | 78.377 |
| Case 6 | 51.700 | 51.174 | 50.789 | 50.744 |
| Case 7 | 86.600 | 86.889 | 87.202 | 87.158 |
| Case 8 | 74.740 | 74.882 | 74.928 | 74.947 |
| Case 9 | 52.190 | 51.576 | 51.260 | 51.127 |
| Paired t-test results | | | | |
| t-statistic (p-value) | | 0.140 (0.893) | -0.64 (0.538) | -1.2 (0.265) |

### 4.2.5 Various factors and levels

To investigate the factors which might affect the clustering performance, we added the variations of error schemes with nine different combinations of cross-correlations and auto-correlations. We considered all available ranges of the levels for the cross-correlation as stated above: weak, moderate, and strong. For the auto-correlation, all the available ranges of the levels are considered also. ARMA (1, 1) which has fixed MA component ($\theta_1 = 0.9$) and three different AR components ($\phi_1 = 0.67$, 0.79, and 0.9) were generated. This study is interested in the short-term time series only which has less than 30 temporal points. For applying to clinical trial cases we eventually applied to, the minimum number of temporal points is six. The middle figure was arbitrary selected as 15 temporal points. We must check if there is any effect on the folding method which is necessary steps to make the multivariate time series available to use the principal component analysis. There are two eligible ways to apply for this study since the observation should be clustered in the end ($I\times(P\times T)$, $I\times(T\times P)$, $I$ = observation, $T$ = temporal points, $P$ = variables). When kernel function was applied to, the RBF parameter, c, should be considered if there is any effect on the clustering performance results. For this simulation, c = 1

and 2 are tried. This is a trial and error procedure to find the most appropriate parameter to get the most successful clustering rate.

## 4.3 Simulation study II

The second simulation study will see if the proposed method can distinguish multivariate time series data into two groups which had two different mean profiles but same error structures. It is mentioned that two different time series can have different mean profiles, two different error structures, or both. Therefore, once we know that given time series data have different mean profiles, then we can conclude that these belong to different groups. However, for multivariate time series cases, there are much more complicated associations such as cross-correlations among the variables, auto-correlations between the temporal points within the variable, or auto-correlations between the temporal points among the variables. These relations can mask true mean profiles so that the time series data can be looked as similar groups. In this simulation study, we attempt to investigate how KMPCA works with those environments.

### 4.3.1 Generating two groups with different mean profiles

Two different mean profiles are generated with the non-linear function presented above. The first group has three different non-linear schemes as the simulation study I had, model 1, 2, and 3. The second group has the same non-linear function with different functional parameters, which are model 4, 5, and 6 as shown in figure 4.2.

For both groups, we used the same factors: three different cross-correlations levels (weak, moderate, and strong) and three different auto-correlations levels ($\phi_1$ = are 0.67, 0.79, and 0.9) with fixed MA component ($\theta_1$ = 0.9). Therefore, there are nine cases of different error structures.

### 4.3.2 Factors to be considered

We are interested in the same factors considered in the first simulation study. They are three levels of the number of temporal points, two different RBF parameter, and two different folding methods. Therefore, a total number of combinations of comparison cases is 108 ($3\times3\times3\times2\times2$). In sum, this simulation study has same conditions as the simulation study I has except that this study uses two different mean profiles.

### 4.3.3 Number of simulation

Unlike the first simulation study, this study does not have a reference group which is consistent for all through the cases. But, we still have a hypothesis test whether a multivariate time series observation belongs to a group or another while it is not possible to compare each case's result with another's result. So, whether a given observation would be from the group 1 or from the group 2 is a binomial distribution ~ b(NSIM, $\pi$) where $\pi$ = P[A certain observation belongs to the group 1] and NSIM is the number of simulation. Still, we do not know true proportion of success, so to determine the size of simulation we set up the margin of error in estimating the proportion of success in advance as 0.05. Therefore, the simulation size, NSIM, is obtained to meet the specified margin of error

by $$NSIM = \hat{P}\cdot(1-\hat{P})\cdot\left(\frac{z_{\alpha/2}}{0.05}\right)^2 = 0.25\left(\frac{1.96}{0.05}\right)^2 \approx 385\,. \qquad (4.19)$$

For the simulation study II, we have two different mean profile groups with nine different combinations of auto-correlations and cross-correlations. Three cases of temporal points, two different folding methods, and two different RBF parameters are still kept for the study. Therefore, a total of 108 cases are generated. Number of simulation is set as 385, and number of each group is 200.

# CHAPTER 5

# SIMULATION RESULTS EVALUATION

The main purpose of this chapter is to present simulation results of clustering multivariate time series data using Kernel Multi-way Principal Component Analysis (KMPCA). The first objective of the simulation study is to test if KMPCA can distinguish a multivariate time series data observation from the one which has different error structure or different mean profiles. In addition, we attempted to know what factors would affect on the clustering performances of KMPCA from the simulation results. For these objectives, we simulated two different cases: first, we generated multivariate time series data with pure error only of which error variance proportion to total variance is 10% and those which have the same mean profiles with various combinations of error schemes such as cross-correlation and auto-correlation. To investigate the factors which can affect on clustering performances, we varied number of temporal points, kernel function parameter, and folding method. The second simulation is for testing if the proposed method would work when the multivariate time series data have the different mean profiles but same combinations of error structures. This study would help us to understand the proposed method better such as when we apply this method and how we can seek better results changing the conditions of the methods.

## 5.1 Simulation study I

The first simulation study was designed for checking if the proposed method can distinguish a tri-variate, short-period time series data from the one which follows same mean profiles but has different error structure. A group of three mean profiles used for both the reference case and comparison cases is borrowed from the previous quality control study (Kim and Adams 2009). The parameters of the formula were manipulated for this research purpose.

The reference group has 10% pure error (N (0, $\sigma^2$)) only whereas the comparison group has various combinations of cross-correlations and auto-correlations added to the error variance structure based on estimated covariance structure. Table 5.1. provides the features of nine cases. To make the number of simulation sample size be larger than necessary and the margin of error be less than required, a total of 385 simulation cases were generated. The number of sample for both groups was set as 200 for the perspective of the efficiency of calculating time after the comparison tests showed no differences as described in previous chapter.

Table 5.1. Nine Schemes of Generating Multivariate Time Series Data

| Case | $\Sigma_E$ | $\phi_1$ |
|---|---|---|
| 1 | Weak | 0.68 |
| 2 | Weak | 0.79 |
| 3 | Weak | 0.9 |
| 4 | Moderate | 0.68 |
| 5 | Moderate | 0.79 |
| 6 | Moderate | 0.9 |
| 7 | Strong | 0.68 |
| 8 | Strong | 0.79 |
| 9 | Strong | 0.9 |

To test if number of temporal points affect on clustering performance, we set temporal points as 30, 15 and six. This study is interested in the short-term time series only which has less than 30 temporal points. Also, we attempted to apply the proposed method to clinical trial cases which is necessary to have more than six time points conventionally. And, we arbitrary selected

15 temporal points to see if there exist any quadratic effects of this factor. To keep nonlinearity in the mean profiles, we used the front parts for six and 15 temporal points' cases.

There are two eligible folding methods ($I\times(P\times T)$, $I\times(T\times P)$, $I$ = observation, $T$ = temporal points, $P$ = variables), and two different kernel function parameters (c = 1 and 2) which had been used from previous research. Therefore, a total of 108 cases were considered to compare the clustering success rates.

Each simulation case has generated 400 observations composed of 200 observations for the reference group and 200 observations for control group. After transforming from three-dimensional data matrix ($I\times P\times T$) to two-dimensional data matrix ($I\times(P\times T)$) or from ($I\times T\times P$) to ($I\times(T\times P)$), we applied to radial basis kernel function with two different parameters, c = 1 or c = 2. We obtained principal component scores by ordinary principal component analysis using this kernel matrix.

We clustered those scores using *k*-means algorithm to limit the potential groups as two. So, we have four possible cases as seen in the table 5.2. Therefore, the total success rate is (A + B)/400. And, a total of 385 iterative simulations provided an average of success rate for comparison between the reference group and one of nine control group cases.

Table 5.2. Four Cases of Clustering Results

| | Clustered Reference Group | Clustered Control Group |
|---|---|---|
| Actual Reference Group | Success (A) | Fail |
| Actual Control Group | Fail | Success (B) |

### 5.1.1 Factor screening

To screen the factors which do not affect on the clustering performance, we used $2^k$ factorial design analysis. Before factor screening, we dropped the error variance proportion to total variance since it depends on auto-correlation as described in the previous chapter. So, to

fulfill this analysis, all five factors, cross-correlation, auto-correlation, RBF parameter (c = 1 and 2), temporal points, and folding method were considered. If the factor has more than two levels, two extreme levels were included in the test, two levels of cross-correlation (Weak and Strong), auto-correlation ($\phi_1$ = 0.67 and 0.9), and temporal points (t = 6 and 30).

In this analysis, each case has only one average success rate. Therefore, to select the factors and interaction terms which have statistically significant effects on the mean clustering success rate, a normal plot of the effects or a half normal plot was presented for the unreplicated factorial design like this. The effects that are negligible are normally distributed, with mean zero and variance $\sigma^2$ and will tend to fall along a straight line on this plot, whereas significant effects will have nonzero means and will not lie along the straight line.

As seen in the figure 5.1, auto-correlation, temporal points, and RBF parameter did not follow the linear line. Especially, auto-correlation has the biggest effect. For the interaction terms, auto-correlation with temporal points and auto-correlation with RBF parameter were out of the line, which means these two interaction terms need to be studied more by detail.

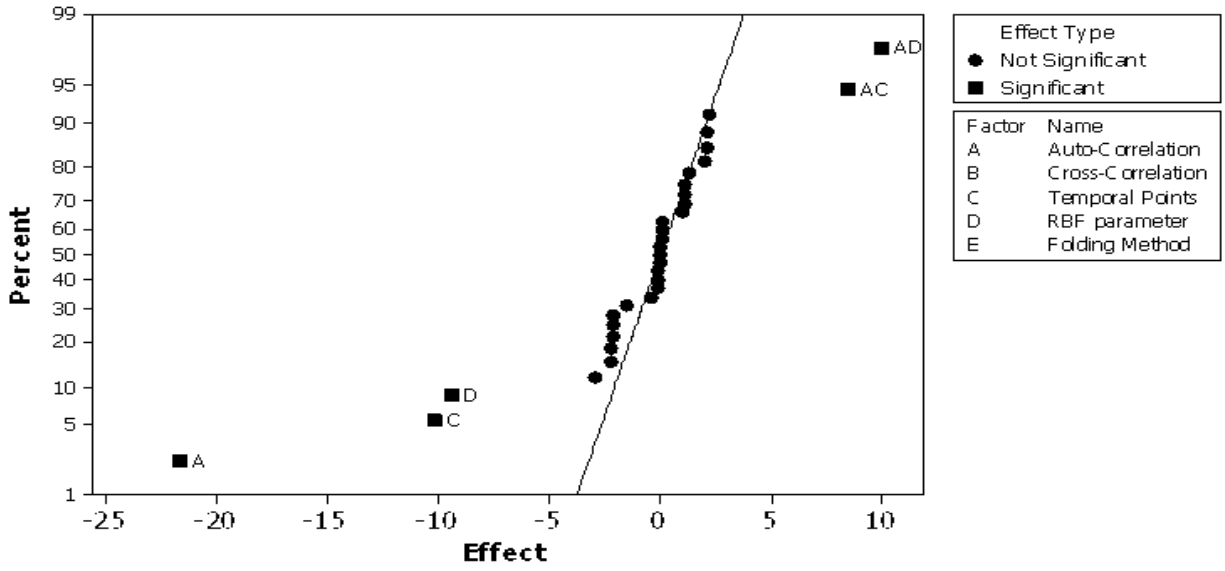


Figure 5.1. Normal Plot of the Effects for Simulation Study I

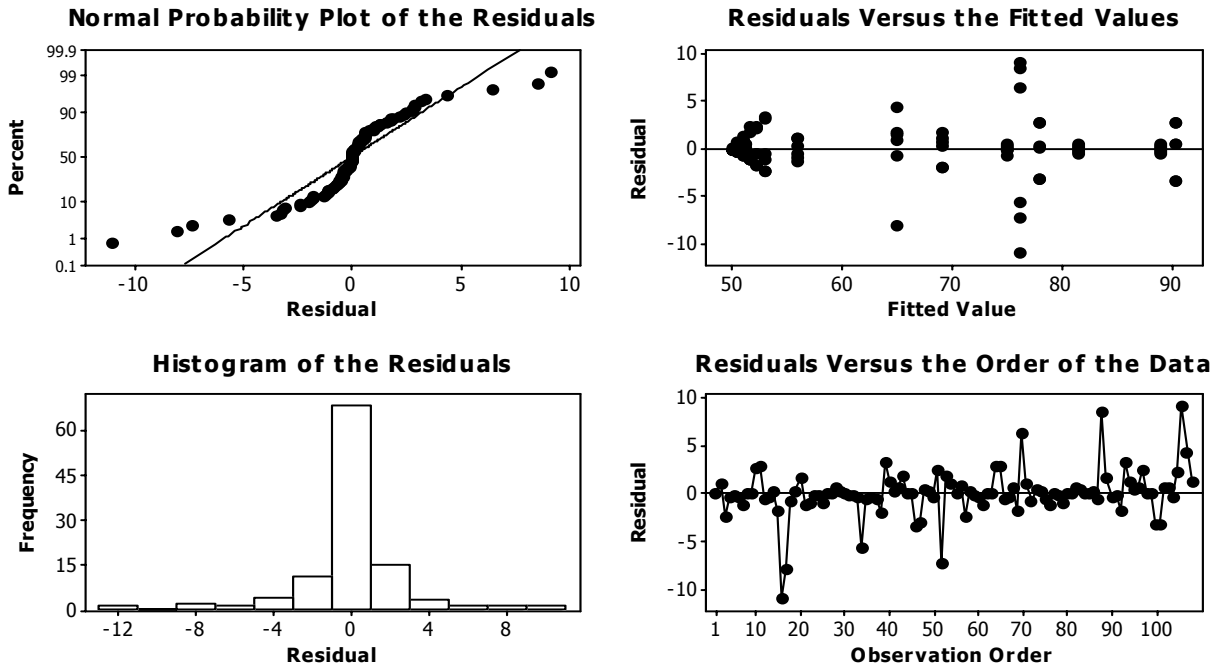


Figure 5.2. Diagnostic Residual Plots for Normal Plot of Simulation Study I

Before the interpretation of the results, the residual analysis to check for model adequacy was done. As seen in the residual plots, the assumptions are roughly met to the satisfaction level; however, it looks like the data set have several unusual observations.

As seen in the ANOVA table, all three factors and interaction terms are statistically significant. These explain almost of all the variances (96.6%) of the data. Using this pre-screening process, those three factors are in the mixed model analysis for further review.

Table 5.3. ANOVA Table of $2^k$ Factorial Design Analysis for Simulation I (Minitab 15.0)

Analysis of Variance for Success_Rate, using Adjusted SS for Tests

| Source | DF | Seq SS | Adj SS | Adj MS | F | P |
|---|---|---|---|---|---|---|
| RBF Parameter | 1 | 5500.9 | 5500.9 | 5500.9 | 749.19 | 0.000 |
| Temporal Points | 2 | 2656.0 | 2656.0 | 1328.0 | 180.87 | 0.000 |
| Auto-corr. | 2 | 8814.7 | 8814.7 | 4407.3 | 600.26 | 0.000 |
| RBF Parameter*Temporal Points | 2 | 948.0 | 948.0 | 474.0 | 64.55 | 0.000 |
| RBF Parameter*Auto-corr. | 2 | 2811.3 | 2811.3 | 1405.6 | 191.44 | 0.000 |
| Temporal Points*Auto-corr. | 4 | 1041.5 | 1041.5 | 260.4 | 35.46 | 0.000 |
| RBF Parameter*Temporal Points* Auto-corr. | 4 | 320.9 | 320.9 | 80.2 | 10.93 | 0.000 |
| Error | 90 | 660.8 | 660.8 | 7.3 | | |
| Total | 107 | 22754.1 | | | | |

S = 2.70969 R-Sq = 97.10% R-Sq(adj) = 96.55%

5.1.2 Simulation results

Main effects

The reference group has pure error with N (0, σ) only with ten percent of the total variance. Therefore, as the auto-correlation was decreasing from 0.9, to 0.8, to 0.68, which means the proportion of error variance to total variance was increasing from 10%, to 20%, to 30%, the success rate of clustering was getting better from 51.25%, 65.72%, to 72.98% as seen in the Figure 5.3.

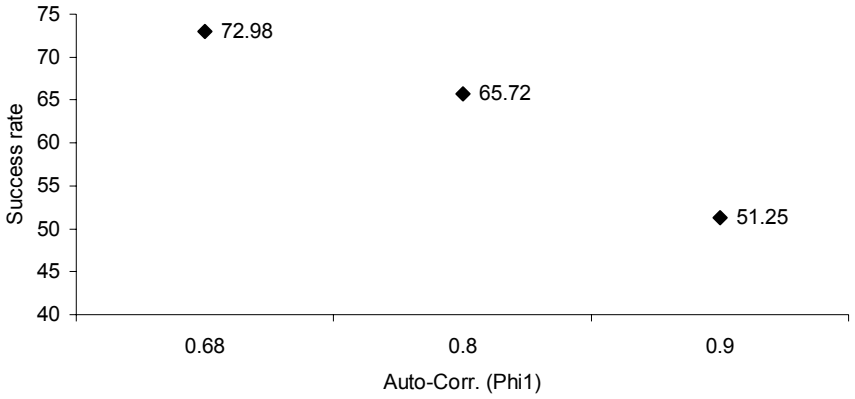


Figure 5.3. The Success Rates by the Different Auto-Correlations for Simulation Study I

As explained in the former chapter, the proportion of error variance to total variance depends on the auto-correlation level. Therefore, when the compared group has the same error variance proportion of total variance with the reference group which is $\phi_1 = 0.9$, the generated data are laying in the same distance range from the mean profiles. In addition, since the cross-correlation does not affect on the performance results, even though the covariance structure was added with various cross-correlations, both groups are similar enough not to tell one specific group from each other. This will explain why the clustering performances increase as auto-correlations decrease which means the error variance proportion to total variance is getting much more different from that of the reference group.

As temporal points were getting smaller, the clustering performance showed much better results. Especially when the number of temporal point is six, the results showed over 90% successes in many cases. It is suspected that an increase in temporal points results in an increase in complexity as well as failure rates.

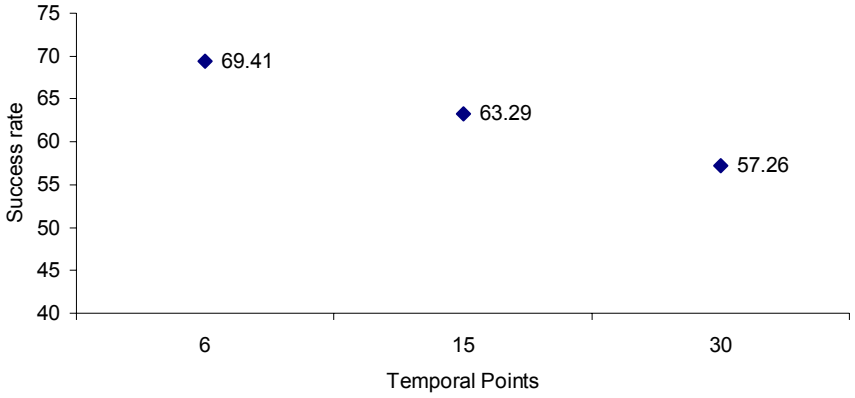


Figure 5.4. The Success Rates by the Different Number of Temporal Points for Simulation Study I

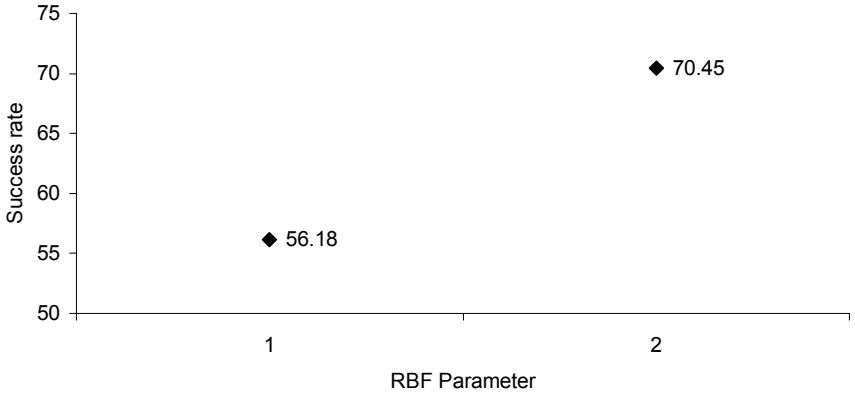


Figure 5.5. The Success Rates by the Different RBF Parameters for Simulation Study I

In MKPCA, the selection of a kernel function is the most important since the degree of capturing nonlinear characteristic of a system is dependent upon it (Lee, et al. 2004). Therefore, selecting the RBF parameter will be decided by the characteristics of the data set. The result statistically proved that the clustering performance differed. It was improved from 56.18% to

70.45%. So, it is suggested that to find an appropriate kernel function with proper parameter is critical in practice.

Interaction Effects

No matter what the RBF parameter was, the clustering success rate was dropping as the auto-correlation increases, which corresponds to the results as main factor effect result. However, the interaction effect between the RBF parameter and the auto-correlation showed statistically significant. As the parameter, c = 1, the rate does not show much difference as the auto-correlation was changing from 0.68 to 0.9. But, when the parameter was two, the success rate dropped significantly from 85% to 51%. From the results, the conclusion is easily drawn that finding the appropriate parameter for given kernel function would be one of the most important procedures to apply the KMPCA to cluster multivariate time series data. We used the parameters, c = 1 or 2, after we took trial and error procedure of preliminary analysis using the same data set. Since we have three variables and short temporal points less than 30, it did not take much time. However, if the number of variables and number of temporal points increase, to find an appropriate parameter will be a burden but a critical step to get more success in clustering. If a clustering result would be very different with historical data or practical wisdom, it is suggested to run more cases with different RBF parameters.

Figure 5.7 showed the clustering success rates considering RBF parameter and number of temporal points changing. When the RBF parameter was eight, the success rate was stable around 73% even though the number of temporal points was changed from six to 15. When c = 1, the success rates dropped drastically to around 50%. When the t = 6, there were not much differences between the results of c = 1 and that of c = 2.

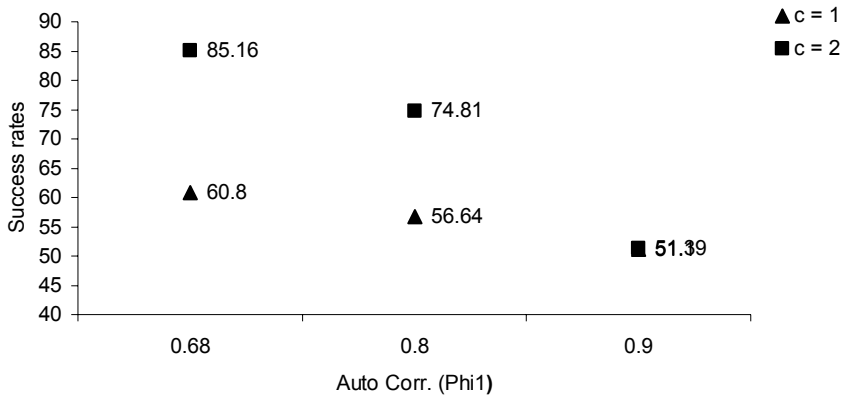


Figure 5.6. The Success Rates by the Different Auto-Correlations and RBF Parameters for Simulation Study I

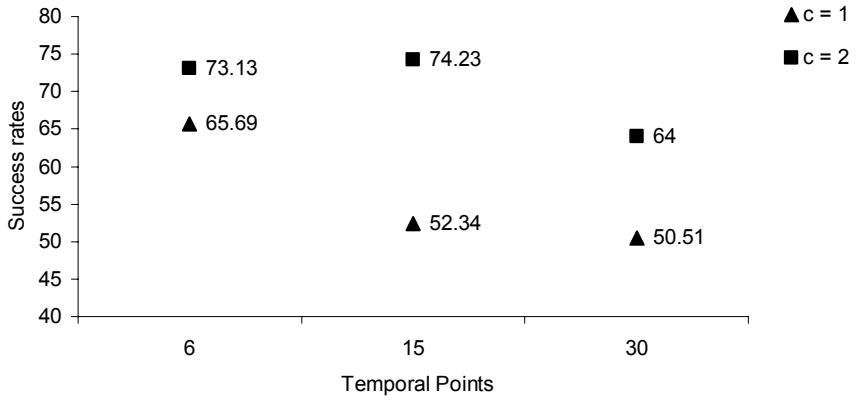


Figure 5.7. The Success Rates by the Different Number of Temporal Points and RBF Parameters for Simulation Study I

These results stressed that the best strategy to get higher clustering success rate is to find an appropriate RBF parameter for a given data structure. Since the data with small number of temporal points may have less complex structure, it seems that the proposed method would be not affected by the kernel function parameter much.

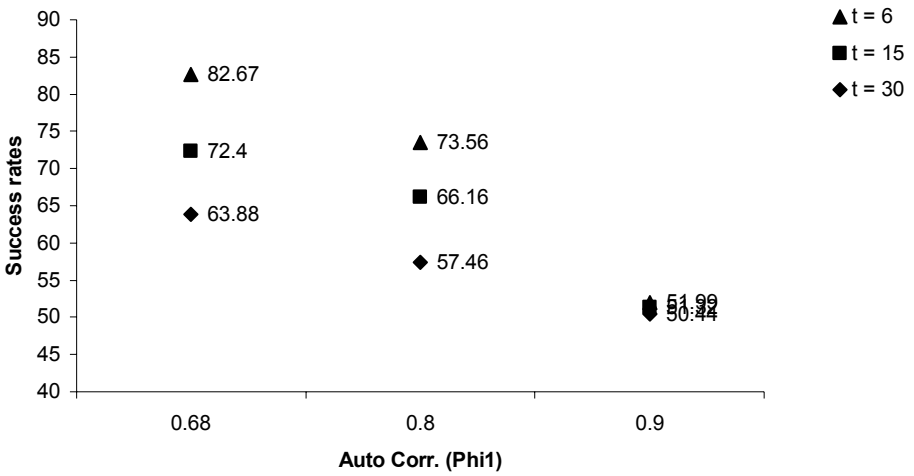


Figure 5.8. The Success Rates by the Different Auto-Correlations and Number of Temporal Points for Simulation Study I

We have already seen from ANOVA results that auto-correlation has statistically significant interaction effects with number of temporal points, which corresponds to figure 5. 8 showed. When $\phi_1 = 0.68$, as the number of temporal points is getting smaller, the success rates are getting higher results. When $\phi_1 = 0.8$, the rates were smaller than before since the complexity of data increase. However, when $\phi_1 = 0.9$, the error variance to total variance is same with that of reference group so that the clustering performances for three different cases of number of temporal points are getting very much close to each other. Compared to other number of temporal points, the simplest data structure (when t = 6) showed better results. So, even if we lost statistical validity of the model with small number of temporal points in time series data, we can obtain reasonable explanation of the complex data using the proposed method.

### 5.1.3 Conclusion of simulation study I

We investigated if the proposed method can cluster two different trivariate time series data generated from the same mean profiles. Auto-correlation has statistically significant effects

on the clustering performance; however, it seems that the differences of success rates resulted from the ratio of error variance to total variance rather than auto-correlation when mean profiles are same. However, it is very hard to know statistically valid auto-correlation coefficients for short-term time series data. We cannot control this factor in practice for real-world data.

The number of temporal points affected clustering success rates; smaller number of temporal points is literally simple so that it results in a higher success rate while bigger number of that is related to complexity of data structure so that it produces poor clustering rates.

Among the effects of main factors, the kernel function parameter is the most critical factor to consider obtaining better performance. It is mentioned that we already screened out other kernel function from preliminary tests. In sum, for a given data, selecting a proper kernel function with appropriate parameter is the most important procedure to increase clustering performance.

From the results of interaction effects, we fortified that an appropriate kernel function with proper parameter should be sought for given data for better clustering performances. Even though other factors were involved, the right function can distinguish the observations from the other group. However, it can take much time and effort. It is suggested to seek alternative ways to find the proper parameter in the future.

## 5.2 Simulation study II

In most real world cases, it could easily expected that two different mean profiles of multivariate time series data were compared and attempted to be distinguished from each other. For this case, we investigated if two different mean profiles can be distinguished from each other by the proposed method when both groups have same error structures in the covariance scheme

with same number of temporal points applying the same folding method and kernel function with same parameter.

This time, two different mean profiles were generated using the non-linear function as used before. The first mean profile has three different non-linear profiles as the simulation study I had, model I, II, and III. The second profile has the same non-linear function with different function parameters as explained in the previous chapter, model IV, V, and VI.

From these two different mean profiles, we estimated covariance structures. We varied cross-correlations and auto-correlations as former study had. Nine different schemes were multiplied to both of the mean profiles. Therefore, we considered three different levels' cross-correlations (Weak, Moderated, and Strong), three different auto-correlations (ARMA (1, 1)) with $\theta_1$=0.9 and $\phi_1$ = 0.67, 0.79, and 0.9).

Other factors were considered: we checked three different levels of number of temporal points (t = 6, 15, and 30), two RBF parameters (c = 1 and 2), and two different folding methods in this simulation study. So, a total of 108 cases were generated and compared.

In here, we do not have a reference group consistent for all through the cases. Therefore, it is unable to compare the results to each other. However, we can still keep a binomial distribution whether an observation is included in group A or not. So, the same number of simulation and the number of observation for each group can be applied. The simulation was iterated 385 times, and the number of observation for each group is fixed as 200 as previous simulation study had.

### 5.2.1 Factor screening

We considered all the same factors with two extreme levels for factor screening as done for the simulation study I. We have cross-correlation (Weak and Strong), auto-correlation ($\phi_1$ = 0.67 and 0.9), RBF parameter (c = 1 and 2), temporal points (t = 6 and 30), and folding method ($I\times(P\times T)$ and $I\times(T\times P)$).

Because of a single replicate of average clustering success rate for each case, normality plot of effects was presented in Figure 5. 10. As seen in the plot, all the factors except folding method have statistically significant effects on the clustering performance since all of the factors are not along the normality line. Also, all the second, third, and fourth order interaction terms showed significant impacts on the success rate. Therefore, all these factors and interaction terms will stay in the final model.

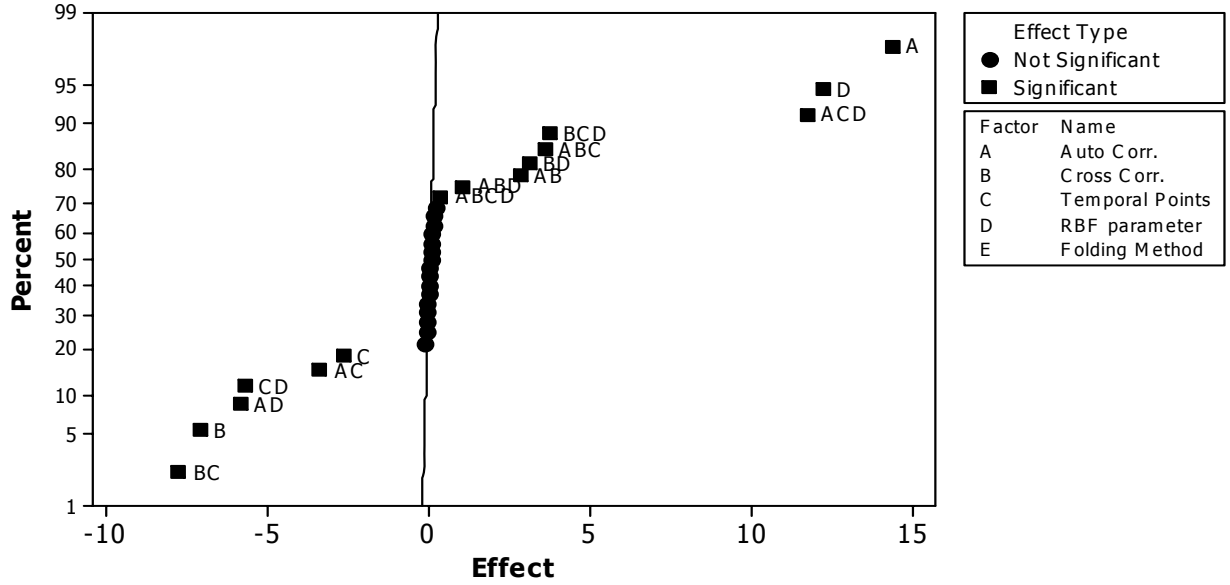


Figure 5.9. Normal Plot of the Effects for Simulation Study II

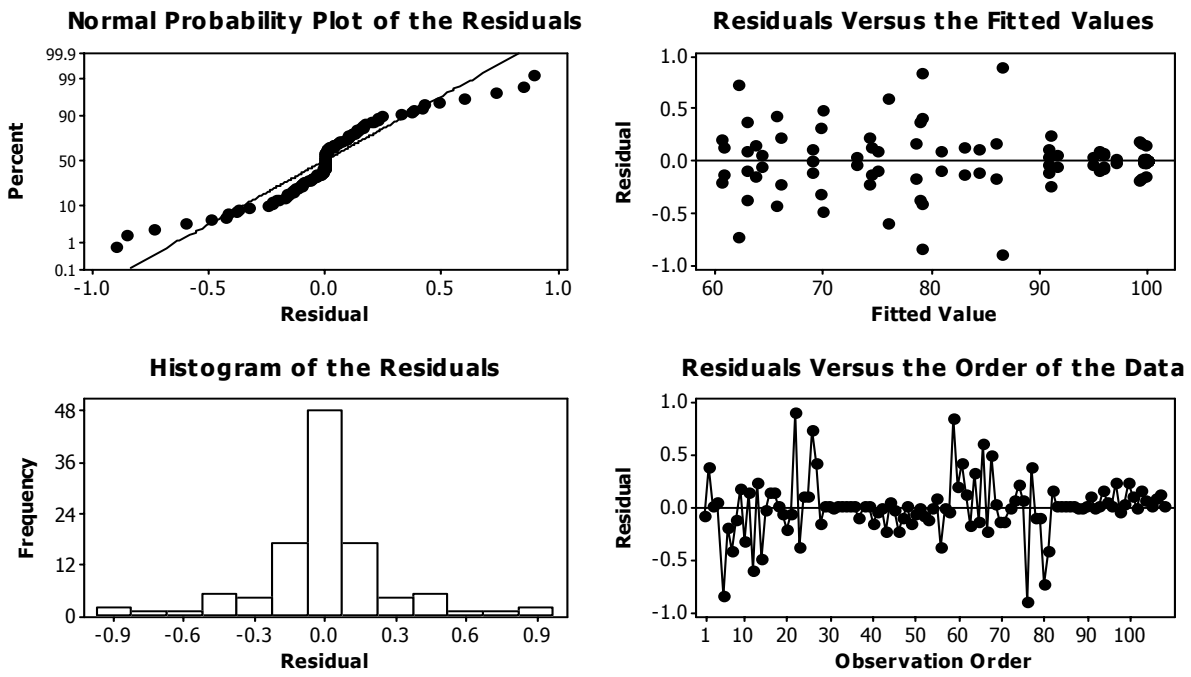


Figure 5.10. Diagnostic Residual Plots for Normal Plot of Simulation Study II

Using the residual analysis, we found the normality plot had still some curvature pattern. And, the independency in order assumption was somewhat violated. However, these did not look problematic. So, the full factorial design model was analyzed with four factors with all the levels considered.

Table 5.4 showed the Minitab results from which we found all the terms have statistically significant effects. All these terms explained almost all the variances.

Table 5.4. ANOVA Table of $2^k$ Factorial Design Analysis for Simulation II (Minitab 15.0)

Analysis of Variance for Success_Rate, using Adjusted SS for Tests

| Source | | Seq SS | Adj SS | Adj MS | F | P |
|---|---|---|---|---|---|---|
| RBF parameter | 1 | 8542.62 | 8542.62 | 8542.62 | 58310.90 | 0.000 |
| Temporal Points | 2 | 2055.91 | 2055.91 | 1027.96 | 7016.70 | 0.000 |
| Cross Corr. | 2 | 600.58 | 600.58 | 300.29 | 2049.73 | 0.000 |
| Auto Corr. | 2 | 3941.55 | 3941.55 | 1970.77 | 13452.27 | 0.000 |
| RBF parameter*Temporal Points | 2 | 307.84 | 307.84 | 153.92 | 1050.63 | 0.000 |
| RBF parameter*Cross Corr. | 2 | 26.26 | 26.26 | 13.13 | 89.64 | 0.000 |
| RBF parameter*Auto Corr. | 2 | 622.40 | 622.40 | 311.20 | 2124.23 | 0.000 |
| Temporal Points*Cross Corr. | 4 | 517.44 | 517.44 | 129.36 | 883.00 | 0.000 |
| Temporal Points*Auto Corr. | 4 | 1245.68 | 1245.68 | 311.42 | 2125.72 | 0.000 |
| Cross Corr.*Auto Corr. | 4 | 119.07 | 119.07 | 29.77 | 203.20 | 0.000 |
| RBF parameter*Temporal Points* Cross Corr. | 4 | 53.34 | 53.34 | 13.33 | 91.02 | 0.000 |

| | | | | | | |
|---|---|---|---|---|---|---|
| RBF parameter*Temporal Points* Auto Corr. | 4 | 2824.47 | 2824.47 | 706.12 | 4819.88 | 0.000 |
| RBF parameter*Cross Corr.*Auto Corr. | 4 | 113.29 | 113.29 | 28.32 | 193.33 | 0.000 |
| Temporal Points*Cross Corr.* Auto Corr. | 8 | 222.23 | 222.23 | 27.78 | 189.61 | 0.000 |
| RBF parameter*Temporal Points* Cross Corr.*Auto Corr. | 8 | 140.54 | 140.54 | 17.57 | 119.91 | 0.000 |
| Error | 54 | 7.91 | 7.91 | 0.15 | | |
| Total | 107 | 21341.13 | | | | |

S = 0.382755 R-Sq = 99.96% R-Sq(adj) = 99.93%

### 5.2.2 Simulation results II

#### Main factors

#### RBF parameter

When the mean structures were different, the RBF parameter explained the most of the variances of the effects. And, when the parameter is 2, the success rate is over the targeted rate, 0.9. Therefore, in practice, it should be the one of the most important tasks to find an appropriate kernel function with proper parameter for a given dataset.

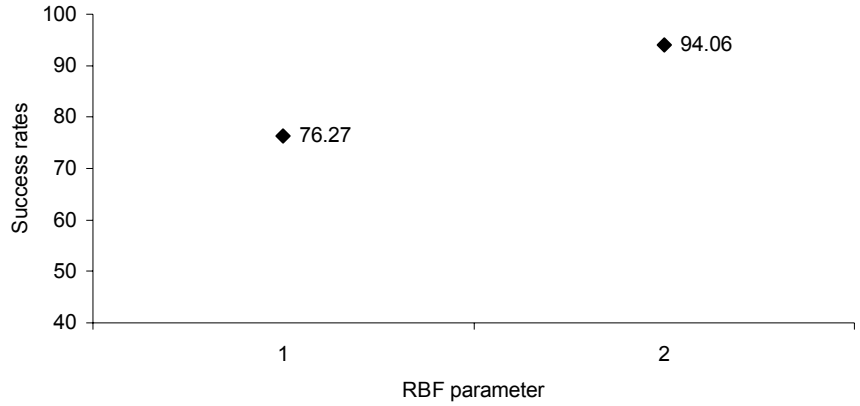


Figure 5.11. The Success Rates by the Different RBF parameters for Simulation Study II

#### Cross-Correlation

As seen in the figure 5.12, as cross-correlation is getting stronger, the clustering success rate decreases, which corresponds to the previous simulation results. Cross-correlations between the time series variables tend to affect the other variables, such as pulling or pushing certain

temporal points of other variables, so that the data has more complexity which would interfere to distinguish the different patterns between the groups.

However, even though cross-correlation is statistically significant as an independent factor, the success rate did not show that it has much impact on the clustering performance. Therefore, it would be better elect the variables which are not so much correlated with other variables in practice, but it will not be an initial job to do.

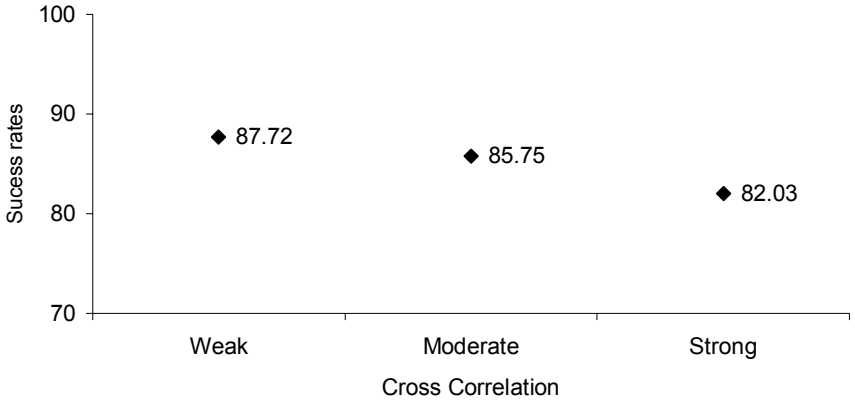


Figure 5.12. The Success Rates by the Different Cross-Correlations for Simulation Study II

Auto-Correlation

Simulation II study gave the error structure with auto-correlation, ARMA (1, 1) with fixed $\theta_1$ = 0.9 and varied $\phi_1$ (0.68, 0.8, and 0.9). When other factors are kept the same, two different mean profiles groups will be distinguished from each other better when auto-correlation is higher. If auto-correlation is higher, the margin of error at each temporal point is smaller, which means data points will be around the mean profiles closely. And, the proposed method detects the difference between the groups mostly depending on not the error structure but the mean structure in this context. That explains that strong auto-correlation had biggest success in clustering.

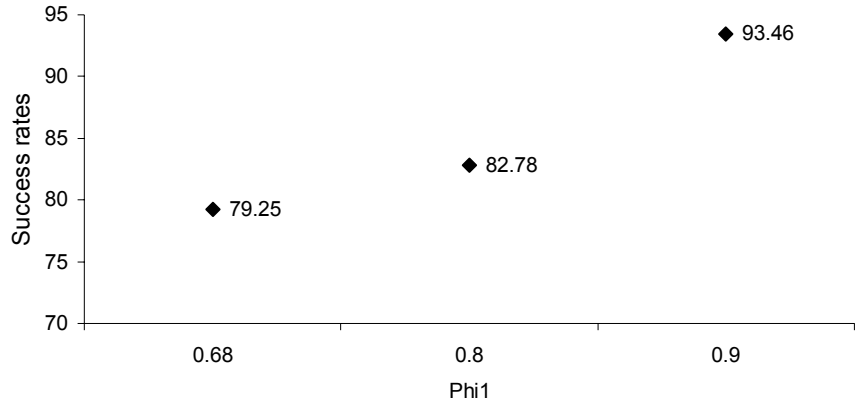


Figure 5.13. The Success Rates by the Different Auto-Correlations for Simulation Study II

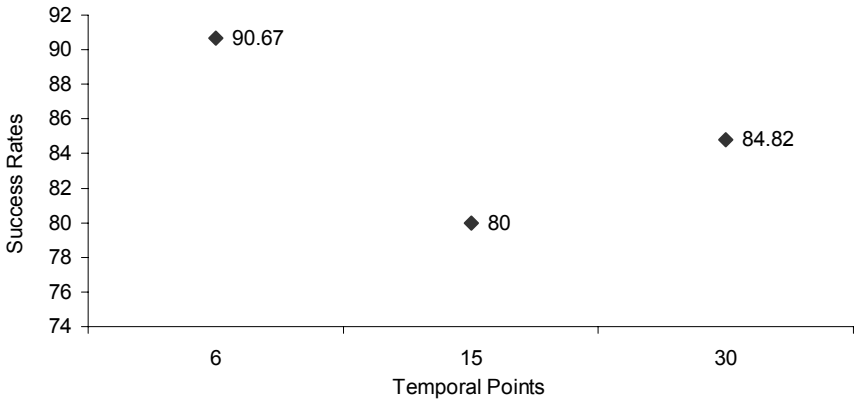


Figure 5.14. The Success Rates by the Different Number of Temporal Points for Simulation Study II

Number of Temporal Points

From the results of the simulation I study, we noticed that number of temporal points may be related with data complexity since shorter temporal points had better success rates. But, figure 5.14 revealed different story. Still the shorter time series data can be easier to cluster with KMPCA; when number of temporal points is 6 has the best clustering success rate, around 91%. But, when t = 30, it had better success rate, around 85% than when t = 15 of which 80%. It may be resulted from the nonlinearity remained after fitting the kernel function. For the decisive

conclusion, more cases of different temporal points should be studied later. So, we will discuss about this more with the results of interaction effects.

Interaction Effects

The number of temporal points showed a U-type change in clustering success rates. When the number of temporal points interacted with the kernel function parameter, the same pattern was shown. When c = 1, the dropping was much bigger than c = 2. It looked like the dropping was from when RBF parameter is two which did not fit to the data structure well. So, the performance does not show consistent results at all. However, whatever the RBF parameter is, the results showed U-shaped changes as the number of temporal point increases.

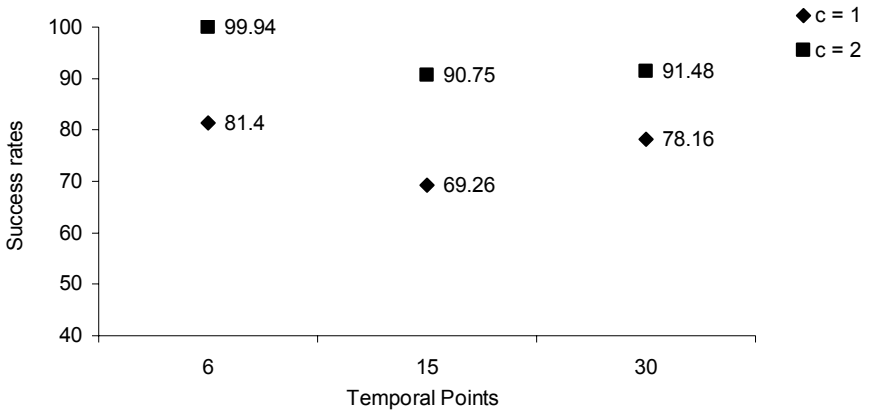


Figure 5.15. The Success Rates by the Different Number of Temporal Points and RBF Parameter for Simulation Study II

Figure 5.16 presented the clustering success rates when number of temporal points and cross-correlations are considered simultaneously. The shorter the temporal point is, the higher success rate is, which is correspond to former results. And, the cross-correlation has the same results as before; t = 15 has the lowest success rates. Therefore, the complexity of data seems not

to be the only reason of differences of success rates. The nonlinearity left in the data can be another reason of these results. This should be examined in future research. Figure 5.17 showed other interaction plots between number of temporal points and auto-correlations. Shorter temporal points and strong auto-correlations had better results as before. And, when t = 15, it had the lowest results. One thing to be noted is that, when $\phi_1 = 0.8$, the success rates were 90.2%, 81.9%, and 76.3% for t = 6, 15, and 30 respectively. For the decisive conclusion, different number of temporal points should be studied later.

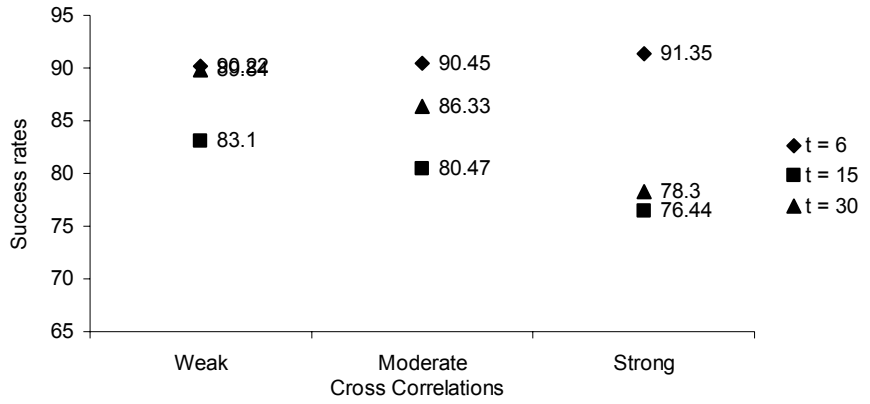


Figure 5.16. The Success Rates by the Different Cross-Correlations and Number of Temporal Points for Simulation Study II

RBF parameter has significant interaction effects with cross correlation, auto correlation, and temporal points. When RBF parameter was involved with cross-correlations, success rate has almost same pattern with that of the factor alone. However, as seen in the next graphs, when c is two, an appropriate parameter in this case study, the performances are mostly consistent, over 90% success rate. When c is one, the rate dropped drastically overall. Therefore, optimal parameter seeking has more weight on than any other task for the successful procedure. Roughly saying, whatever the parameter is, the performance pattern followed the involved interactive factor's independent pattern.

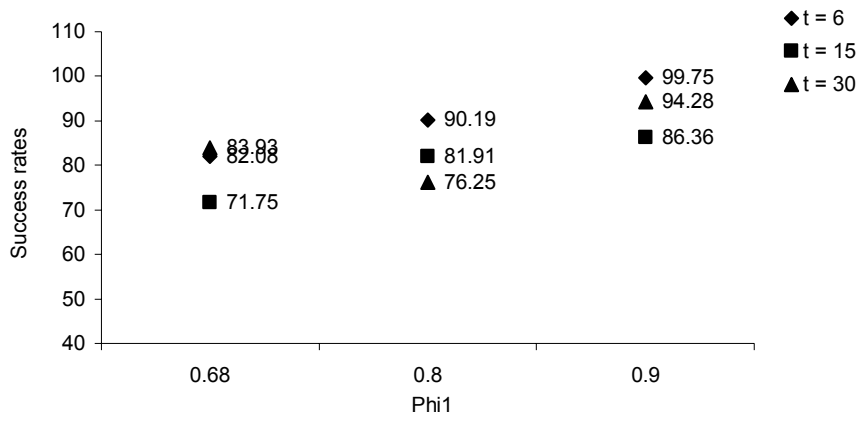


Figure 5.17. The Success Rates by the Different Auto-Correlations and Number of Temporal Points for Simulation Study II

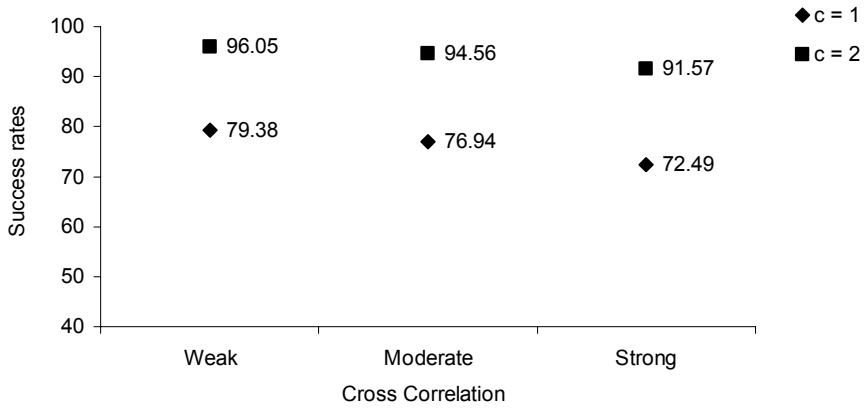


Figure 5.18. The Success Rates by the Different Cross-Correlations and RBF Parameters for Simulation Study II

When the auto correlation was interacted with cross correlation, strong auto-correlation and weak cross-correlation will have the most successful clustering results. Especially, when the data has strong auto-correlation, no matter what the extent of cross-correlation, success rates are over 90%. For the practitioners who tried to cluster short time series data, it would be helpful to get more variables which would have strong auto correlation with weak cross correlation for obtaining better clustering performance.

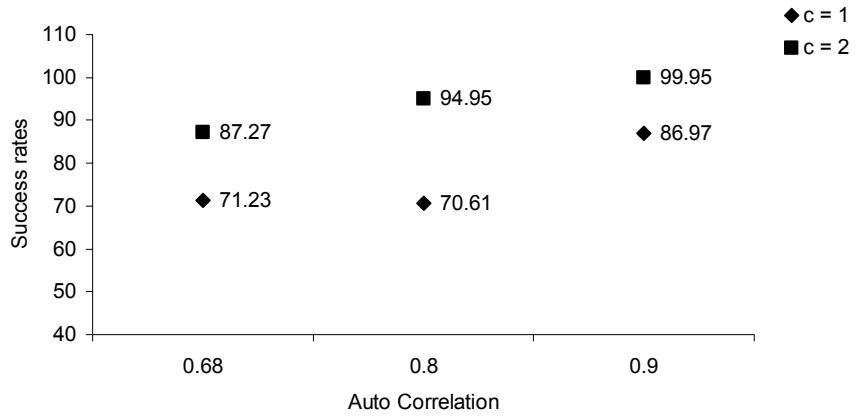


Figure 5.19. The Success Rates by the Different Auto-Correlations and RBF Parameters for Simulation Study II

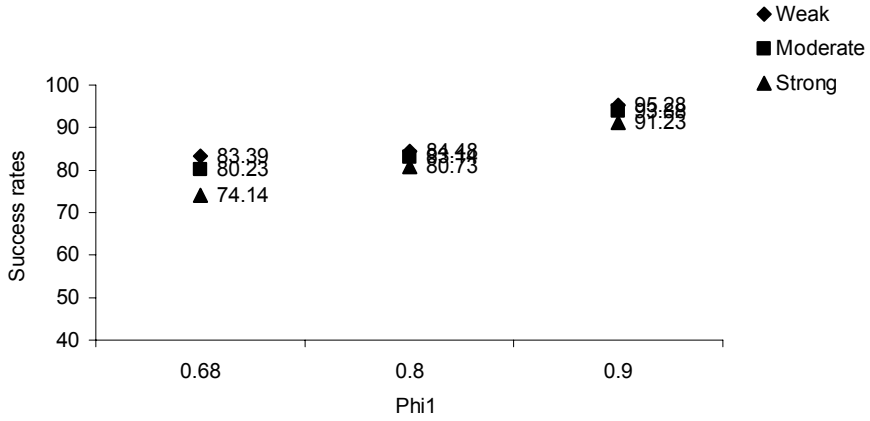


Figure 5.20. The Success Rates by the Different Cross-Correlations and Auto-Correlations for Simulation Study II

### 5.2.3 Conclusion of simulation study II

Simulation study II had two different mean profiles groups clustered into two groups when they had the same error structures. Auto-correlation, cross-correlation, RBF parameter, and number of temporal points have statistically significant effects on the clustering performance. And, all interaction effects related with those main effects were also statistically significant. Folding methods do not affect on the clustering performance.

Like the results of simulation study I, RBF had the biggest effects on the success rates. In practice, it is difficult to find an appropriate kernel function with proper parameter when there are many time series variables and observations. So, it is suggested that the researcher take preliminary steps to screen the variables with a test data set which is a part of given data.

Strong auto-correlation makes the data in the small margin of error for each temporal point, so it has the best clustering success rate. Strong cross-correlation looks like to pull or push other variables' observation from their own mean profiles so that it can mask true mean profiles, even though the effects are not large. Therefore, it is suggested to try to avoid strong correlated variables in the data matrix. From these results, we can notice that mean structure difference only would be easier to cluster than the error structure difference mixed with mean structure. And, error structure can affect on the mean structure with strong cross-correlation and weak auto-correlation.

Shorter temporal point will help cluster much better in this simulation case. However, the success rates dropped when t = 15 and went up again when t = 30. It is suspected that this results from (1) poor performances due to wrong RBF parameter (c =1), (2) remained nonlinearity of the mean profiles around t = 10, or (3) unidentified reasons. For the decisive conclusion, it should be studied more in the future.

# CHAPTER 6

# CASE STUDY: CLUSTERING THE DEPRESSED AMONG THE CARDIA (CORONARY ARTERY RISK DEVELOPMENT IN YOUNG ADULTS) STUDY WITH OBESITY RELATED VARIABLES USING KMPCA

We investigated if the proposed method, KMPCA will cluster a real world data into a couple of groups. We used the Coronary Artery Risk Development in Young Adults (CARDIA) study sample. Applying to KMPCA with radial basis function, we attempted to cluster CARDIA cohorts into two groups, mentally depressed group at 2005 (year 20) based on the Center for Epidemiologic Studies Depression Scale (CES-D) and non-depressed group with obesity related variables measured six times at 1985 (Year 0), 1987 (Year 2), 1990 (Year 5), 1992 (Year 7), 1995 (Year 10), and 2000 (Year 15) exams.

## 6.1 Introduction

Obesity and depression are two major diseases which are associated with many other health problems such as hypertension, dyslipidemia, diabetes mellitus, coronary heart disease, stroke, myocardial infarction, and heart failure in patients with systolic hypertension, low bone mineral density, and increased mortality. Both diseases share common health complications but there are inconsistent findings concerning the relationship between obesity and depression. We suspect the reason for inconsistency is that most researchers have been seeking the results with cross sectional studies which might overlook the time effects of the variables. In this work with the cohorts of the CARDIA study we used the KMPCA to examine the relations between body

mass index (BMI), as a proxy for obesity, and depression when considering several related longitudinal variables simultaneously.

The Coronary Artery Risk Development in Young Adults (CARDIA) Study is a study examining how heart disease develops in adults. It began in 1986 with a group of 5,115 black and white men and women aged 18-30 years. The participants were selected so that there would be approximately the same number of people in subgroups of race, gender, education (high school or less and more than high school) and age (18-24 and 25-30) in Birmingham, AL; Chicago, IL; Minneapolis, MN; and Oakland, CA. These same participants were asked to participate in follow-up examinations during 1987-1988 (Year 2), 1990-1991 (Year 5), 1992-1993 (Year 7), 1995-1996 (Year 10), 2000-2001 (Year 15), and 2005-2006 (Year 20).

In this study, we investigated the association between the obesity or obesity-related variables and depression in general population considering time effect. For this objective, we gathered obesity related variables measured at all six exams without missing values. With this data matrix, we applied the proposed method, KMPCA with radial basis function and attempted to cluster the depressed among CARDIA cohorts based on CES-D score surveyed at year 20.

In addition, we explored if there is gender or racial effect to expect in the depression group. Our objectives for these findings are (1) to improve our understanding of the association between obesity and depression and (2) to provide a template for estimating depression in the population at large.

## 6.2 Literature Review

Many researches have addressed the relationship among obesity, general psychopathology, and depression in particular for decades. However, there are inconsistent findings concerning the relationship between obesity and depression.

Some studies concluded that there was no association between obesity and depression. Friedman and Brownell reviewed the cross-sectional association between obesity status and depression. Their meta-analysis of four previous studies revealed a small, nonsignificant association (Friedman and Brownell 1995). However, this finding was from study heterogeneity since many studies used different concepts of depression, obesity, and so on for their own researches. Faith et al. reported that there was no significant relationship among men, whereas greater BMI was associated with significantly greater neuroticism which is correlated with depression in women. The authors tested the association using multiple regression models between BMI and psychological variables in a British population-based sample (Faith, et al. 2001).

Others argued that there is inverse association between obesity and depression. Palinkas et al. argued that heavier people were less depressed (Palinkas, et al. 1996). They examined 2,245 noninstitutionalized men and women aged 50 to 89 years living in California. They found that the prevalence of Beck Depression Inventory scores was inversely associated with body weight in men, but not in women using logistic regression analyses.

In contrast, some studies reported that there is positive association between these two variables. Goodman et al. concluded that depressed adolescents are at increased risk for the development and persistence of obesity during adolescence (Goodman, et al. 2002). They surveyed 9,374 adolescents in grades 7 through 12 about depressed mood and calculated BMI from self-reported weight and height at 1995 for baseline. And, they surveyed them again one year later. Logistic models were used for analysis. They reported that baseline obesity did not predict follow-up depression.

Simon et al. used nationally representative sample of over 6,000 US adults (Simon, et al. 2006). The authors used logistic models to figure out the association between obesity and psychiatric disorders such as major depression, bipolar disorder, and panic disorder. They found that obesity was associated with significant increases in lifetime diagnosis of major depression. But, they found not any gender difference but racial and education effects.

Methodological differences across studies have contributed to these inconsistent observations. Friedman and Brownell already noted that most population-based studies had not defined depression according to established psychiatric diagnostic criteria. Obesity also had defined differently in several studies. Onyike et al. divided the cohorts from the Third National Health and Nutrition Examination Survey (1988-1994) into several groups according to obesity severity. They used logistic regression to check if there is any association between obesity and depression. The authors found that obesity is associated with depression mainly among persons with severe obesity.

Stunkard et al. approached to find reasons of discrepancies using the moderators and mediators. They introduced several potential moderators such as severity of depression, severity of obesity, gender, socioeconomic status, and gene-environment interactions. And, as potential mediators, eating and physical activity, disordered eating, and stress were listed. However, the authors still suggested that prospective studies be necessary to clarify the obesity-depression association and that better understanding of genes that promote both depression and obesity (Stunkard, et al. 2003).

Roberts et al. reported that when obesity and depression were examined prospectively, controlling for other variables such as age, sex, education, marital status, social isolation and social support, chronic medical conditions, functional impairment, life events, and financial

strain, obesity in 1994 predicted depression in 1995 using logistic regression (Roberts, et al. 2000). However, the prospective multivariate analyses were not significant. So, this study just provided overall suggestion of an association between obesity and depression.

Most of these studies are cross-sectional analyses. Very few studies have examined the prospective development of obesity—depression correlations. Even though some considered time effect in the model, their measure was limited two times, baseline and follow-up. So, the need for new, multivariate prospective studies cannot be overstated (Faith, et al. 2002).

For several decades, the association between obesity and depression was sought by extensive researches. Many of them concluded that there is any kind of association between obesity and depression. These days, novel approaches were attempted. Some suggested extra association with other variables between these two major variables. Others addressed that new variable like gene expression would solve the discrepancies. Some others tried to add time effects into models to explain the association, but still not ripe in this branch. Therefore, to overcome this inconsistency, a study considering time effects with more information from obesity related variables such as BMI, socioeconomic, disease symptoms correlated with obesity, and so on, might be another way of challenge.

## 6.3 Method

### Sample

In 1985-1986, 5,115 adults aged 18-30 years were recruited into CARDIA at four sites: Birmingham, AL; Chicago, IL; Minneapolis, MN; and Oakland, CA. The sampling strategy resulted in a population-based cohort that was balanced by race (52% black), sex (55% female), and education (40% with ≤ 12 years of education) both overall and within each clinical center (Friedman et al. 1988). Follow-up examinations were conducted in 1987-1988 (year 2), 1990-

1991 (year 5), 1992-1993 (year 7), 1995-1996 (year 10), 2000-2001 (year 15), and 2005-2006 (year 20). A majority of the group has been examined at each of the follow-up examinations (90%, 86%, 81%, 79%, 74%, and 72%, respectively). Therefore, we left the participants who took all seven examinations.

However, the variables we are interested in this study do not have all the participants or all the year exams. So, we first selected the variables measured every exam and relative to this study based on the previous research.

Variables

Obesity was represented by Body Mass Index (BMI) which was calculated as [(weight (lbs)/2.2)/ (height (cm)/100)$^2$]. Most researches considered physical activity and eating habit are closely related to obesity. However, CARDIA has eating habit variables for only year 0, 7, and 20. It was addressed that age at measurement can be used as a confounder (Onyike, et al. 2003).

Socioeconomic variables can be other confounders for this study. Many studies considered some variables covariates such as income, education, full-time job, marital status, employment, and so on. We chose available variables among CARDIA variables measured at all the exams.

CES-D is one of the most common screening tests for helping an individual to determine his or her depression quotient. The self-test measures depressive feelings and behaviors during the past week. At year 7, 10, 15, and 20, this survey was done for CARDIA study. A score over 16 indicates depression.

To obtain more information, we added the obesity related disease symptoms such as hypertension and cholesterol. So, we collected diastolic blood pressure (DBP), systolic blood pressure (SBP), total cholesterol, total high-density lipoprotein, and total low-density lipoprotein.

Some studies argued that addictive material uses be regarded as covariates in the model. Conventionally alcohol drinking, cigarette smoking, and illegal drug use like marijuana were used in previous researches.

Table 6.1 presented the concepts related to this study and CARDIA variables available. For extended studies, Gender and Race were included for checking the effects.

Table 6.1. Available CARDIA Variables Related to the Clustering Study

| **Concept** | **CARDIA Variables** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | **CARDIA Exams** | | | | | | |
| | | | Year 0 | Year 2 | Year 5 | Year 7 | Year 10 | Year 15 | Year 20 |
| Obesity | Body Mass Index | Height | o | o | o | o | o | o | o |
| | | Weight | o | o | o | o | o | o | o |
| | Physical Activity | Total Score | o | o | o | o | o | o | o |
| Depression | CES-D score | | x | x | o | x | o | o | o |
| Age | Exam Age | | o | o | o | o | o | o | o |
| Socioeconomic | Working full-time | | o | o | o | o | o | o | o |
| | No work | | o | o | o | o | o | o | o |
| | Education | | o | o | o | o | o | o | o |
| | Marital Status | | o | o | o | o | o | o | o |
| Obesity related variables | Blood pressure | Diastolic BP | o | o | o | o | o | o | o |
| | | Systolic BP | o | o | o | o | o | o | o |
| | Cholesterol | Total Cholesterol | o | o | o | o | o | o | o |
| | | Total HDL | o | o | o | o | o | o | o |
| | | Total LDL | o | o | o | o | o | o | o |
| Addictive habits | Smoking | | o | o | o | o | o | o | o |
| | Alcohol Drinking | | o | o | o | o | o | o | o |
| | Illegal Drug | | o | o | o | o | o | o | o |
| Race | Race | | | | | | | | o |
| Gender | Sex | | | | | | | | o |
| Exclusion | Pregnancy | | o | o | o | o | o | o | o |
| | Mental Disorder | | x | x | o | x | o | o | o |

## 6.4 Analysis

### 6.4.1 Factors

Mean profiles: We learned whether the mean structures of two groups are different or not has drawn different clustering results. The same mean profiles case showed that the different auto-correlation, temporal points, and RBF parameter would affect on clustering performance significantly while the different mean profiles case had cross correlation effects as well as three other factors as above.

It was reported that CARDIA study was designed to have the cohorts who have very similar characteristics. As seen in figure 6. 1, there is not any discrepancy between the depressed and non-depressed based on BMI pattern from year 0 to year 15.

However, there is no statistical test to check whether these mean profiles are similar or not so that we should consider all the effects except for the folding method which did not show any significant result from both simulation studies.

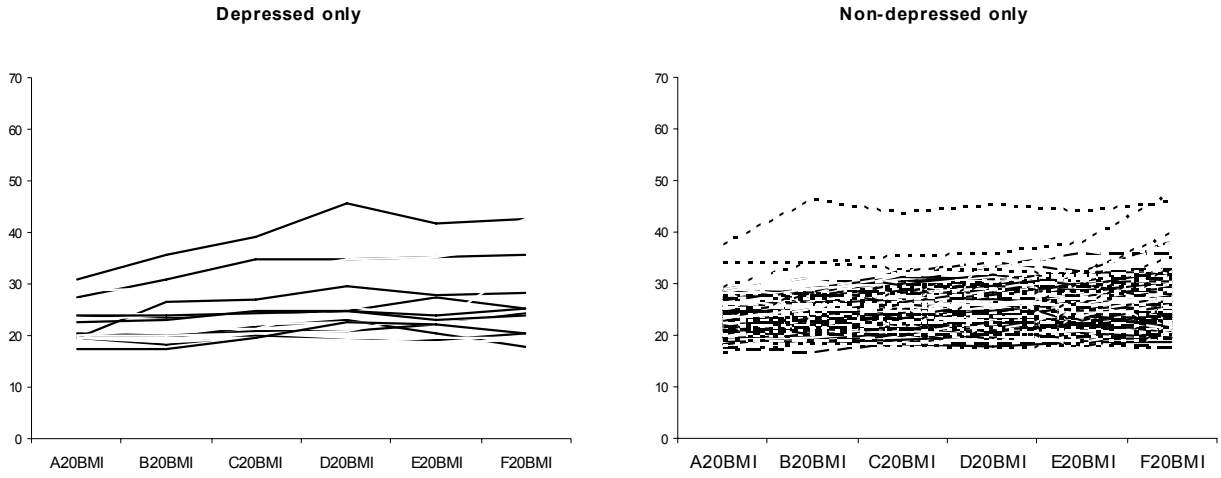


Figure 6.1. Trend Plots of BMI variable by the Depressed Group and Non-depressed Group

Temporal points: CARDIA study measurement was done seven times since 1985. So, temporal points have six points from exam 0, 2, 5, 7, 10, to 15. Based on the simulation results,

we found that smaller points would have better clustering performance using KMPCA since the data complexity is less than that longer temporal point data has. However, simulation studies showed that temporal points can have interaction effects with RBF parameter or with auto-correlations.

Cross-correlation: When the mean profiles are same, we found that the cross-correlation does not have statistically significant effect on the clustering results from the simulation study I. In other hand, simulation II study showed that strong cross-correlation can lessen the clustering performance. Cross-correlation has significant interaction effects with temporal points or with auto-correlations. When the number of temporal point is small such as this case, weak cross-correlation will help to obtain better clustering results.

Figure 6.2 is the cross-correlation matrix of CARDIA year 0 exam variables in which we are interested. From the matrix, we can see that most variables do not have any association. And, even if some variables have statistically significant correlation with each other, the correlation coefficients are as less than 0.3 at most. In the simulation study, we consider these cases as weak cross-correlation which helps to get better results thanks to avoiding data complexity.

Auto-correlation: We have only six temporal points in data matrix so that we cannot have statistically valid test to check auto-correlation coefficients. As stated, we cannot control auto-correlation in real world case. However, this study attempted to show the proposed method can work with short-term time series data with these obstacles.

Kernel function (Radial Basis Function: RBF) parameter, c: As found from the simulation studies, the kernel function parameter, c, should be sought based on a given data. Therefore, we tried various parameters, c from 2 to 9 for better clustering results. As for the

folding method, $I\times(P\times T)$ was only used since we found that it does not have any significant effects statistically.

Pearson Correlation Coefficients, N = 200
Prob > |r| under H0: Rho=0

| BMI | AGE | Physical Activity | Working Fulltime | Unemployment | Education | Marital Status | Smoking | Illegal drug | Drinking | SBP | DBP | Cholesterol | HDL | LDL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.08623 | -0.05929 | 0.08872 | -0.0455 | -0.0881 | -0.13943 | 0.04208 | 0.07708 | 0.06526 | 0.20752 | -0.01277 | 0.17677 | -0.23057 | 0.25669 |
| | 0.2247 | 0.4043 | 0.2116 | 0.5223 | 0.2148 | 0.0489 | 0.5541 | 0.278 | 0.3586 | 0.0032 | 0.8576 | 0.0123 | 0.001 | 0.0002 |
| | 1 | -0.16313 | 0.3675 | -0.19668 | 0.30966 | -0.32947 | 0.11798 | 0.11897 | 0.02533 | -0.03633 | 0.03855 | 0.13025 | 0.03534 | 0.12765 |
| | | 0.021 | <.0001 | 0.0052 | <.0001 | <.0001 | 0.0961 | 0.0934 | 0.7218 | 0.6095 | 0.5878 | 0.066 | 0.6193 | 0.0717 |
| | | 1 | 0.11443 | -0.00056 | -0.04147 | 0.19537 | -0.12189 | 0.10829 | 0.15789 | 0.13756 | 0.00508 | -0.07617 | 0.10998 | -0.11647 |
| | | | 0.1066 | 0.9938 | 0.5599 | 0.0056 | 0.0855 | 0.1269 | 0.0256 | 0.0521 | 0.9431 | 0.2837 | 0.1211 | 0.1005 |
| | | | 1 | -0.20153 | 0.16518 | -0.15362 | -0.0831 | 0.06082 | -0.01972 | 0.05662 | 0.04732 | 0.05899 | 0.0244 | 0.04648 |
| | | | | 0.0042 | 0.0194 | 0.0299 | 0.242 | 0.3923 | 0.7816 | 0.4258 | 0.5058 | 0.4067 | 0.7316 | 0.5134 |
| | | | | 1 | -0.09878 | 0.12356 | 0.10165 | 0.08479 | 0.21847 | 0.00335 | -0.00494 | -0.00166 | 0.11817 | -0.05426 |
| | | | | | 0.164 | 0.0813 | 0.1521 | 0.2326 | 0.0019 | 0.9624 | 0.9446 | 0.9814 | 0.0956 | 0.4454 |
| | | | | | 1 | 0.0178 | -0.27616 | -0.03602 | 0.04695 | -0.09458 | -0.09993 | 0.05074 | 0.117 | 0.026 |
| | | | | | | 0.8025 | <.0001 | 0.6126 | 0.5091 | 0.1828 | 0.1592 | 0.4755 | 0.099 | 0.7148 |
| | | | | | | 1 | -0.07604 | 0.02074 | 0.10971 | 0.09148 | 0.08742 | -0.05567 | 0.08082 | -0.07182 |
| | | | | | | | 0.2845 | 0.7707 | 0.122 | 0.1976 | 0.2184 | 0.4336 | 0.2553 | 0.3122 |
| | | | | | | | 1 | 0.23638 | 0.12139 | -0.04716 | -0.16231 | 0.0507 | -0.09094 | 0.0782 |
| | | | | | | | | 0.0008 | 0.0869 | 0.5073 | 0.0217 | 0.4759 | 0.2003 | 0.271 |
| | | | | | | | | 1 | 0.26881 | 0.13053 | -0.01042 | -0.03743 | -0.07002 | -0.03015 |
| | | | | | | | | | 0.0001 | 0.0654 | 0.8836 | 0.5988 | 0.3245 | 0.6717 |
| | | | | | | | | | 1 | 0.13002 | -0.01256 | -0.07058 | 0.09986 | -0.1399 |
| | | | | | | | | | | 0.0665 | 0.8599 | 0.3206 | 0.1595 | 0.0482 |
| | | | | | | | | | | 1 | 0.60717 | 0.04052 | -0.12098 | 0.07157 |
| | | | | | | | | | | | <.0001 | 0.5689 | 0.0879 | 0.3139 |
| | | | | | | | | | | | 1 | 0.07144 | 0.01293 | 0.05471 |
| | | | | | | | | | | | | 0.3148 | 0.8558 | 0.4416 |
| | | | | | | | | | | | | 1 | 0.23171 | 0.90338 |
| | | | | | | | | | | | | | 0.001 | <.0001 |
| | | | | | | | | | | | | | 1 | -0.15542 |
| | | | | | | | | | | | | | | 0.028 |
| | | | | | | | | | | | | | | 1 |

Figure 6.2. Half Diagonal Matrix of Cross-Correlation among Variables at Year 0

After obtaining the principal components from the analysis, we used k-means algorithm as a clustering method as used in the simulation studies. Total success rate and depressed cohorts prediction rate were compared to get better results among different kernel function parameters. In here, we should weight more on the depressed prediction rate than the total success rate since it is much more critical to find the depressed in advance and thus to help the practitioners to prevent people from the depression symptoms and eventually keep them from the extreme events

such as suicide. Therefore, we were trying to find cases of which both prediction rates are over at least 50% and to pick a case of which the depressed prediction rate is the highest among them.

## 6.5 Results

### 6.5.1 Total cohorts

Among the cohorts (n = 1,178) who did not have any missing values for the variables, we excluded some who did not satisfy the criterion. Pregnancy can affect the weight information. For the exact weight measure, pregnancy people were screened before the examination. A patient with a history of medication for a mental disorder was excluded from the total sample. Age will not change, but it stayed in the data matrix since it might affect the obesity related disease symptoms. Therefore, a total of 229 participants were excluded, and 949 were left for the total sample.

The ratio between the depressed at year 20 and the non-depressed were around 12.5%. We randomly selected 200 cohorts from the data with the same rate of the depressed at year 20. Therefore, non-depressed group has 175 cohorts and the depressed are 25. Using mean and standard deviation of each temporal point, we standardized each variable when the variable is a continuous variable. With this randomly selected and standardized data set, we checked how well the proposed method would work to cluster them between the depressed and non-depressed.

We put only the obesity variable, BMI in the data matrix to see if it could predict the depressed people. Even if we changed the RBF parameter, the depressed prediction rate would not be over 50% which would be better possibility than arbitrary choice could have. The obese variable only could not predict the depression symptoms even if we considered it in the longitudinal context.

After we added obesity related variables such as physical activity scores, we got increased depression prediction rate (52%), but the total prediction rate dropped drastically (53.5%). Blood pressure related variables have been good indicators for obesity prevention research as well as depression prediction studies. We put standardized average systolic blood pressure and average diastolic blood pressure in the data matrix. This time, we got much improved results as seen in the table 6. 2 (total prediction rate: 60.5%, depression prediction rate: 60%).

Adding cholesterol related variables to the data sets did not help to increase success rates for both total and depression prediction. However, as we put more information into total data matrix, so the both prediction rates increase quickly whereas the kernel function parameter increases for better performance results. This means that as the data structure is getting more complexity, it is necessary to have higher order of dimension.

After more trials, we approached different direction as we added socioeconomic variables into the basic variables such as BMI and age. Both prediction rates were over 60% when the parameter was around five. In here, we tried different combinations of socioeconomic variables since they are all categorical variables and do not have much fluctuation exam by exam. And, we found the education did not show any difference in predicting year 20 depression symptoms. Without education, depression prediction rate soared to over 70% even though total prediction rate dropped to around 53%.

If we considered more information, the performances are much more stable, total prediction rate is around over 65% with depression prediction rate is around 60% or so when the parameter is six or bigger. Table 6.2 presented the clustering results of various models we tried. Over 50% success rates were bolded.

Table 6.2. Total and Depression Prediction Rates for Various Models and Different Kernel Parameters Using Total Cohorts (n = 949)

| | | Kernel function parameter, c | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | | c = 2 | c = 3 | c = 4 | c = 5 | c = 6 | c = 7 | c = 8 | c = 9 |
| Model 1: BMI, Age only | Total prediction rate | 0.56 | 0.625 | 0.71 | 0.79 | 0.845 | 0.855 | 0.865 | 0.865 |
| | Depression prediction rate | 0.32 | 0.24 | 0.2 | 0.16 | 0.16 | 0.08 | 0.08 | 0.08 |
| Model 2: Model 1 + Physical Activity score | Total prediction rate | **0.535** | 0.57 | 0.65 | 0.725 | 0.8 | 0.835 | 0.84 | 0.865 |
| | Depression prediction rate | **0.52** | 0.36 | 0.28 | 0.2 | 0.2 | 0.2 | 0.08 | 0 |
| Model 3: Model 2 + Blood pressure | Total prediction rate | 0.645 | 0.505 | **0.605** | 0.67 | 0.77 | 0.795 | 0.81 | 0.83 |
| | Depression prediction rate | 0.16 | 0.36 | **0.6** | 0.36 | 0.28 | 0.24 | 0.12 | 0.12 |
| Model 4: Model 3 + Cholesterol | Total prediction rate | 0.815 | 0.725 | 0.48 | **0.585** | 0.685 | 0.755 | 0.785 | 0.795 |
| | Depression prediction rate | 0.16 | 0.12 | 0.24 | **0.6** | 0.44 | 0.36 | 0.36 | 0.2 |
| Model 5: Model 2 + Socioeconomic | Total prediction rate | 0.86 | 0.6 | 0.62 | 0.52 | **0.54** | **0.61** | **0.55** | **0.555** |
| | Depression prediction rate | 0 | 0.28 | 0.32 | 0.28 | **0.64** | **0.56** | **0.64** | **0.6** |
| Model 6: Model 5 + Addictive material uses | Total prediction rate | 0.865 | 0.77 | 0.645 | 0.54 | **0.565** | **0.64** | **0.605** | **0.575** |
| | Depression prediction rate | 0 | 0.08 | 0.32 | 0.24 | **0.6** | **0.56** | **0.56** | **0.64** |
| Model 7: Model 6 without Education | Total prediction rate | 0.81 | 0.705 | 0.605 | **0.525** | 0.705 | 0.76 | 0.835 | 0.85 |
| | Depression prediction rate | 0.04 | 0.16 | 0.28 | **0.72** | 0.4 | 0.24 | 0.16 | 0.08 |
| Model 8: Model 4 + Socioeconomic | Total prediction rate | 0.875 | 0.82 | 0.655 | 0.65 | 0.535 | **0.555** | 0.545 | 0.545 |
| | Depression prediction rate | 0.04 | 0.04 | 0.12 | 0.2 | 0.4 | **0.64** | 0.32 | 0.6 |
| Model 9: All variables considered | Total prediction rate | 0.875 | 0.78 | 0.695 | 0.61 | 0.58 | **0.545** | **0.605** | **0.675** |
| | Depression prediction rate | 0.04 | 0.08 | 0.08 | 0.2 | 0.2 | **0.64** | **0.6** | **0.56** |

### 6.5.2 Gender differences

We divided the cohorts (n = 949) into male (n = 485) and female (n = 464). Male group has 13% of the depressed (CES- D ≥ 16) at year 20 and female group does 12%. Therefore, after standardized at each temporal point for continuous variables, the cohorts were randomly selected according to those ratios.

Table 6.3. Total and Depression Prediction Rates for Various Models and Different Kernel Parameters Using Male Group Only (n = 485)

| | | Kernel function parameter, c | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | | c = 2 | c = 3 | c = 4 | c = 5 | c = 6 | c = 7 | c = 8 | c = 9 |
| Model1: BMI, Age, Physical Activity score, Blood Pressure, Cholesterol, Socioeconomic | Total prediction rate | 0.825 | 0.785 | 0.690 | 0.580 | **0.505** | **0.525** | 0.555 | 0.625 |
| | Depression prediction rate | 0.115 | 0.115 | 0.231 | 0.231 | **0.538** | **0.500** | 0.462 | 0.385 |
| Model2: BMI, Age, Physical Activity | Total prediction | 0.840 | 0.825 | 0.680 | 0.545 | **0.530** | 0.545 | 0.585 | 0.600 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| score, Socioeconomic, Addictive Material Use | rate | | | | | | | | |
| | Depression prediction rate | 0.000 | 0.000 | 0.192 | 0.346 | **0.500** | 0.462 | 0.462 | 0.385 |
| Model 3: Model 2 without Education | Total prediction rate | 0.850 | 0.520 | **0.545** | **0.515** | **0.535** | **0.530** | 0.710 | 0.745 |
| | Depression prediction rate | 0.115 | 0.423 | **0.538** | **0.577** | **0.577** | **0.577** | 0.192 | 0.154 |
| Model 4: Model 2 without Education and Working full-time | Total prediction rate | 0.825 | 0.615 | 0.510 | **0.515** | **0.530** | **0.530** | **0.545** | 0.745 |
| | Depression prediction rate | 0.115 | 0.385 | 0.423 | **0.577** | **0.577** | **0.577** | **0.577** | 0.115 |

Unlike the total case, male case did not have many cases of which total and depression prediction rates were over 50%. Table 6.2 presented the models which predicted the depressed at year 20 well. Even though several models had over 50% success rates, the biggest success rate was under 60%. Istavan et al. reported in their study that relative body weight was weakly related to elevated depression scores in women but not men (Istavan, et al. 1992). And, some other researchers followed same results.

For women group, the results were not so much different than the male group had. Even though we had more cases which satisfied the condition that both prediction rates were over 50%. However, the success rates were merely over 50%. Without any statistical tests, we cannot conclude that there is no gender effect between obesity and depression. However, these results showed that dividing sample data by gender would not help to distinguish the depressed from the non-depressed considering obesity related variables as these models had.

Table 6.4. Total and Depression Prediction Rates for Various Models and Different Kernel Parameters Using Female Group Only (n = 464)

| | | Kernel function parameter, c | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | | c = 2 | c = 3 | c = 4 | c = 5 | c = 6 | c = 7 | c = 8 | c = 9 |
| Model 1: BMI, Age only | Total prediction rate | **0.510** | 0.685 | 0.755 | 0.805 | 0.800 | 0.835 | 0.865 | 0.865 |
| | Depression prediction rate | **0.583** | 0.333 | 0.292 | 0.250 | 0.208 | 0.208 | 0.125 | 0.125 |
| Model 2: Model 1 + Physical Activity score | Total prediction rate | **0.600** | 0.590 | 0.690 | 0.725 | 0.785 | 0.810 | 0.830 | 0.840 |
| | Depression prediction rate | **0.500** | 0.375 | 0.292 | 0.292 | 0.208 | 0.208 | 0.125 | 0.125 |
| Model 3: Model 2 + Cholesterol | Total prediction rate | 0.850 | 0.735 | **0.575** | 0.525 | 0.645 | 0.685 | 0.720 | 0.730 |
| | Depression prediction rate | 0.125 | 0.208 | **0.500** | 0.417 | 0.333 | 0.292 | 0.208 | 0.208 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model 4: Model 2 + Socioeconomic | Total prediction rate | 0.810 | 0.780 | 0.685 | 0.535 | **0.545** | 0.570 | 0.565 | **0.545** |
| | Depression prediction rate | 0.167 | 0.208 | 0.333 | 0.375 | **0.542** | 0.458 | 0.458 | **0.500** |
| Model 5: Model 4 + Addictive material uses | Total prediction rate | 0.860 | 0.800 | 0.710 | 0.555 | **0.555** | **0.550** | **0.565** | **0.575** |
| | Depression prediction rate | 0.000 | 0.000 | 0.208 | 0.333 | **0.542** | **0.542** | **0.500** | **0.500** |

### 6.5.3 Race differences

This time, the cohorts (n = 949) were divided into African-American (n = 289) and Caucasian (n = 660). African-American group had 19% of the depressed at year 20 and other group did 9.5%. For each group, 200 randomly selected participants with those ratios were set after standardized at each temporal point for continuous variables.

African-American group results showed much better prediction rates in several models. A couple of models showed that total prediction rates are over 60% and the depression prediction rates over 70%. Most of the models with RBF parameter around 5 or 6 had over 60% depression prediction rates. Even the basic information model such as BMI and age only had 66% depression prediction rate when the parameter was 2. One thing to note is that the models which have addictive material uses variables had over 70% or more depression prediction rates.

Table 6.5. Total and Depression Prediction Rates for Various Models and Different Kernel Parameters Using African American Group Only (n = 485)

| | | Kernel function parameter, c | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | | c = 2 | c = 3 | c = 4 | c = 5 | c = 6 | c = 7 | c = 8 | c = 9 |
| Model 1: BMI, Age only | Total prediction rate | **0.555** | 0.595 | 0.67 | 0.735 | 0.75 | 0.775 | 0.775 | 0.775 |
| | Depression prediction rate | **0.658** | 0.289 | 0.132 | 0.053 | 0.053 | 0 | 0 | 0 |
| Model 2: Model 1 + Physical Activity score | Total prediction rate | **0.615** | **0.55** | 0.58 | 0.615 | 0.705 | 0.735 | 0.745 | 0.745 |
| | Depression prediction rate | **0.5** | **0.632** | 0.289 | 0.132 | 0.053 | 0.026 | 0 | 0 |
| Model 3: Model 2 + Blood pressure + Cholesterol | Total prediction rate | 0.79 | 0.67 | 0.555 | **0.56** | 0.62 | 0.7 | 0.73 | 0.735 |
| | Depression prediction rate | 0.079 | 0.237 | 0.342 | **0.605** | 0.368 | 0.263 | 0.211 | 0.158 |
| Model 4: Model 3 + Socioeconomic | Total prediction rate | 0.8 | 0.72 | 0.705 | **0.645** | **0.58** | **0.535** | 0.625 | 0.675 |
| | Depression prediction rate | 0 | 0.211 | 0.342 | **0.526** | **0.579** | **0.711** | 0.289 | 0.237 |
| Model 5: Model 4 + Addictive material uses | Total prediction rate | 0.78 | 0.76 | 0.65 | **0.51** | **0.57** | 0.58 | **0.59** | 0.645 |
| | Depression prediction rate | 0 | 0 | 0.079 | **0.737** | **0.632** | 0.474 | **0.5** | 0.447 |
| Model 6: Model 2 + | Total prediction rate | 0.795 | 0.725 | 0.68 | **0.63** | **0.545** | 0.57 | 0.65 | 0.67 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Socioeconomic | Depression prediction rate | 0.026 | 0.289 | 0.474 | **0.579** | **0.711** | 0.211 | 0.105 | 0.105 |
| Model 7: Model 6 + Addictive material uses | Total prediction rate | 0.79 | 0.67 | **0.56** | **0.59** | **0.58** | **0.53** | 0.655 | 0.535 |
| | Depression prediction rate | 0 | 0.368 | **0.763** | **0.658** | **0.632** | **0.658** | 0.421 | 0.158 |
| Model 8: Model 7 without Education | Total prediction rate | 0.78 | 0.755 | 0.6 | **0.52** | **0.62** | 0.715 | 0.765 | 0.78 |
| | Depression prediction rate | 0.105 | 0.132 | 0.211 | **0.684** | **0.5** | 0.237 | 0.211 | 0.184 |
| Model 9: Model 7 without difficulty of paying | Total prediction rate | 0.795 | 0.74 | **0.615** | **0.61** | **0.6** | **0.535** | 0.66 | 0.69 |
| | Depression prediction rate | 0 | 0.132 | **0.658** | **0.605** | **0.605** | **0.658** | 0.421 | 0.132 |

Caucasian group also predicted the depressed at year 20 in some models. Overall, many models did not satisfy both prediction rates were over 50%. However, addictive material uses variables were included into the model, the depression prediction rates soared up to 84%. Surprisingly, we dropped education variable from socioeconomic variables, the rate reached close to 95%. Therefore, for the further studies of obesity-depression association, it would be better to consider racial effects.

Table 6.6. Total and Depression Prediction Rates for Various Models and Different Kernel Parameters Using Caucasian Group Only (n = 660)

| | | Kernel function parameter, c | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | | c = 2 | c = 3 | c = 4 | c = 5 | c = 6 | c = 7 | c = 8 | c = 9 |
| Model 1: BMI, Age, Physical Activity score | Total prediction rate | 0.610 | **0.515** | 0.645 | 0.710 | 0.790 | 0.820 | 0.840 | 0.840 |
| | Depression prediction rate | 0.474 | **0.579** | 0.421 | 0.263 | 0.053 | 0.000 | 0.000 | 0.000 |
| Model 2: Model 1 + BP, Cholesterol, Socioeconomic | Total prediction rate | 0.895 | 0.785 | 0.700 | 0.620 | **0.565** | 0.510 | 0.660 | 0.710 |
| | Depression prediction rate | 0.000 | 0.263 | 0.158 | 0.421 | **0.579** | 0.316 | 0.263 | 0.158 |
| Model 3: Model 2 + Addictive material uses | Total prediction rate | 0.875 | 0.780 | 0.695 | 0.610 | 0.580 | **0.545** | **0.605** | **0.675** |
| | Depression prediction rate | 0.053 | 0.105 | 0.105 | 0.263 | 0.263 | **0.842** | **0.789** | **0.737** |
| Model 4: Model 1 + Socioeconomic + Addictive material uses | Total prediction rate | 0.885 | 0.660 | 0.535 | **0.535** | 0.525 | 0.515 | 0.670 | 0.610 |
| | Depression prediction rate | 0.000 | 0.158 | 0.474 | **0.526** | 0.474 | 0.421 | 0.421 | 0.211 |
| Model 5: Model 4 without Education | Total prediction rate | 0.855 | 0.840 | 0.605 | **0.525** | **0.705** | 0.760 | 0.835 | 0.850 |
| | Depression prediction rate | 0.105 | 0.211 | 0.368 | **0.947** | **0.526** | 0.316 | 0.211 | 0.105 |

### 6.5.4 Results Summary

From the results of total cohort sample, we found that obesity may not directly relate to depression. It looks like obesity only cannot explain depression even though historical

information was considered. As this issue has been controversial for several decades among extended researches, we did not find simple relationship between those variables.

Once we included more information into data matrix, prediction rates increased. But, it was hard to find a good model with low parameters. It would take more time and efforts for compensation of better model. Also, without prior information or an expert's opinion, it will be a challenge to decide which model would be the most successful clustering rates.

Physical activity score and socioeconomic variables seem to help to cluster the depressed among the sample. But, education did not help much in contrast to the suggestion from previous research. Addictive materials uses variables were helpful to obtain higher success rates.

There seemed no gender effect based on the results of this study. Both prediction rates for men and women had seldom over 50% success rates which are lower than those of total sample. Women group showed simple model with BMI and age only explained depression. Men group had to have more information to do so.

We suggest considering race effect for better model to enlighten the association between obesity and depression. Divided into African-American and Caucasian group, some models had promising success rates; 76% for African-American and 95% for Caucasian. Addictive material uses variables helped considerably to explain of high depression possibility. Yet, education was not helpful again.

## 6.6 Conclusion

This study attempted to cluster selected CARDIA sample into the depressed at year 20 or non-depressed group based on the obesity related variables from year 0 to year 15 applying the proposed method, KMPCA.

Depression and obesity are increasingly prevalent and associated with various health complications. Therefore, to enlighten the association between these two variables and thus to expect the depression symptoms and prevent/treat in advance is more than helpful to health care practitioners as well as potential depression patients.

For decades, researches have been trying to figure out the relationship with diverse approaches, but it is still controversial. Thus, this study applied KMPCA to the obesity related data matrix considering time effect to cluster the depressed near future.

From the analyses, we found that obesity did not predict depression simply. As previous researches reported, obesity related variables, especially addictive material use variables, help to increase total and depression prediction rates.

In contrast to former founding, we did not have gender effect but racial effect. After dividing the sample into African-American and Caucasian groups, depression prediction rates soared to 76% and 95% respectively.

Clustering is rudimentary analysis so that it cannot have decisive results. Principal component analysis serves as intermediate steps in much larger investigations. Therefore, this study cannot provide a final statistically significant conclusion at all. And, it is impossible to compare these results with those of previous research. But, we can provide some inspiration of the association between obesity and depression to specialists and practitioners in this field.

Without collaboration with experts in the related fields, it costs considerable time and efforts to find better model with appropriate kernel function with proper parameter. It might be fruitful to divide a given data into test and validating data to check the model obtained from KMPCA.

The proposed method is constrained to not use missing values since missing values can affect mean structure which is critical to cluster. Therefore, it is suggested to search more related variables available to increase success rates. Or, to find a reasonable way to treat missing values to get more valid observations is another one.

# CHAPTER 7

# CONCLUSION AND FUTURE RESEARCH

## 7.1 Conclusion

Clustering multivariate time series data has been a challenging task for researchers since temporal data has multiple dimensions to consider. The main character of multivariate time series data is that variable profiles are highly correlated with each other. Also measurements over time for each variable profile are highly correlated.

Even though there has been extensive research to analyze time series data in diverse area such as economics, statistics, process control theory, signal processing, and so on, not many had been for clustering multivariate time series data. According to Liao's survey study, three categorized approaches are addressed: raw-data-based, feature-extraction-based, and model-based approaches. Since this study is dealing with short temporal multivariate time series data without prior information, one of feature-extraction-based methods, clustering with KMPCA can be an alternative method.

Principal component analysis is a feature extract method. Multi-way principal component analysis is an extension of PCA to handle data in three dimensional arrays. However, MPCA is a linear algorithm. Therefore, the data with higher order correlations and nonlinearity as multivariate time series has cannot be explained by MPCA. Kernel is a sort of function to help an object to be divided into a certain group. Using several functions such as inner product we can

transform original data into feature space where we can handle much higher dimension. And, kernel variant MPCA can deal with nonlinearity of the data as we have for this study.

We investigated if the proposed method would perform clustering multivariate time series data and provide the guidelines for selecting appropriate clustering schemes to apply KMPCA to. Detailed procedures of generating nonlinear, multivariate time series data are illustrated in Chapter 4.

At first, based on a nonlinear function, six different mean profiles were generated and divided into two groups. There were two simulation cases: same mean profiles groups with different error structure and two different mean profiles groups with same error structure. Except for using same mean profiles or not, both studies have generated same schemes of error structures, three different auto-correlations and three different cross-correlations were considered. Other factors such as three different numbers of temporal points, two eligible folding methods, and two kernel function, radial basis function, parameters were investigated as well. A total of 108 cases were generated.

Chapter 5 presents the results of two simulation studies. Founding from the simulation studies as follows:

1. When mean structures are same, auto-correlation, number of temporal point, and kernel function parameter have statistically significant effects on clustering performance. Second and third order interaction effects with each other of those factors also have effects on the clustering success rates.

2. Among the effects of main factors, kernel function parameter is the most critical factor to consider obtaining better performance. These correspond to the results of interaction effects. Even though other factors were involved, the right function can distinguish the

observations from the other group. However, it can take much time and efforts without prior information or experts' opinions. In practice, it is not an easy work to find an appropriate kernel function with proper parameter when there are many time series variables and observations without prior information. So, it is suggested preliminary steps to screen the variables with a test data set which is a part of given data.

3. Another important factor is the difference of the proportion of error variance to total variance, which means auto-correlation should be distinctively different. However, it is very hard to know statistically valid auto-correlation coefficients with short-term time series data.

4. It looks easier to cluster mean structure difference only than the error structure difference mixed with mean structure. Error structure can affect on the mean structure with strong cross-correlation and weak auto-correlation.

5. When mean structures are different between the groups to cluster with same error structure, auto-correlation, cross-correlation, kernel function parameter, and number of temporal point are statistically significant effects on clustering. Like the results of simulation study I, RBF had the biggest effects on the success rates.

6. Strong auto-correlation makes the data in the small margin of error for each temporal point, so it has the best clustering success rate. Strong cross-correlation looks like to pull or push other variables' observation from their own mean profiles so that it can mask true mean profiles, even though the effects are not much big. Therefore, it is suggested to try to avoid strong correlated variables into the data matrix.

7. Shorter temporal point will help cluster much better than before. However, when t = 15, success rates dropped. It is suspected from poor performances due to wrong RBF parameter (c =1), remained nonlinearity of the mean profiles around t = 10, or unidentified reasons. For the decisive conclusion, it should be studied more in the future.

Chapter 6 presents a real world data case. It was investigated if the proposed method, KMPCA will cluster a real world data into a couple of groups. The Coronary Artery Risk Development in Young Adults (CARDIA) study sample was used. This study attempted to cluster randomly selected CARDIA sample into the depressed at year 20 or non-depressed group based on the obesity related variables for six repeated measures from year 0 to year 15 applying the proposed method, KMPCA. For decades, researches have been trying to figure out the relationship between obesity and depression with diverse approaches, but it is still controversial. Therefore, this study attempted to contribute to this issue and to check if there is gender or racial effects.

The data seems to have similar mean structures, strongly auto-correlated, weakly cross-correlated, and very short temporal points, t = 6. Various radial basis function parameters were tried, and 5 or 6 would work the best in most models. The proposed method is constrained to not use missing values since missing values can affect mean structure which is critical to cluster. Therefore, it is suggested to search more related variables available to increase success rates or to find a reasonable way to treat missing values to get more valid observations. Followings are founding from the study:

1. From the analyses, we found that obesity did not predict depression simply. As previous researches reported, obesity related variables, especially addictive material use variables, help to increase total and depression prediction rates.
2. We found racial effect. After dividing the sample into African-American and Caucasian groups, depression prediction rates soared to 76% and 95% respectively.
3. Clustering and principal component analyses are not final steps for a final conclusion. Therefore, this study cannot provide a final statistically significant conclusion at all. Therefore, we can only provide some inspiration of the association between obesity and depression to specialists in this field.
4. Without collaboration with experts in the related fields, it costs considerable time and efforts to find better model with appropriate kernel function with proper parameter. It might be fruitful to divide a given data into test and validating data to check the model obtained from KMPCA.

## 7.2 Future research

The ideas of future research regarding to improve the proposed method, KMPCA, and limitations of this study are follows:

1) When number of observation, n, is big, considerable time and computing capacity is required to calculate kernel matrix of which dimension is n×n. We selected random sample from total cohorts available. However, it will not easy to do so when there are many variables and lots of observations. Collaborations with other fields such as data mining can provide better solutions for this issue in near future.
2) As indicated in the conclusion section, the performance of the KMPCA is strongly depending on kernel function and parameter. However, there is not much research for

determining optimal parameter on a given data characteristics. It is necessary to find an appropriate kernel function for multivariate time series according to different characteristics for further studies.

3) In real world, time series data can have lots of missing values. Missing values can affect mean structure which is critical to cluster. Therefore, it is suggested to find a reasonable way to treat missing values to get more valid observations.

4) Shorter temporal point will help cluster much better than before. However, when t = 15, success rates dropped. It is suspected from poor performances due to wrong RBF parameter (c =1), inherited nonlinearity of the mean profiles around t = 10, or unidentified reasons. For the decisive conclusion, it should be studied more in the future.

While the clustering multivariate time series is very difficult, research in this area is quite important. But, it cannot overwhelm the necessity of finding abundant information from given data.

# REFERENCES


Agrawal, R., Faloutsos, C., and Swami, A. (1993), "Efficient Similarity Search in Sequence Databases," *International Conference on Foundations of Data Organization*.

Agrawal, R., Lin, K.-I., Sawhney, H. S., and Shim, K. (1995), "Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases," in *21th International Conference on Very Large Databases*, Zurich, Switzerland.

Anderson, R. J., Clouse, R. E., Freedland, K. E., and Lustman, P. J. (2001), " The Prevalence of Cormorbid Depression in Adults With Diabetes-A meta-analysis," *Diabetes Care*, 24, 1069-1078.

Baragona, R. (2001), "A Simulation Study on Clustering Time Series with Metaheusitic Methods," *Quaderni di Statistica*, 3, 1-26.

Box, G. E. P., and Jenkins, G. (1976), *Time Series Analysis: Forecasting and Control*, Holden-Day.

Chan, K.-p., and Fu, A. W.-c. (1999), "Efficient Time Series Matching by Wavelets," in *15th IEEE International Conference on Data Engineering*, Sydney, Australia, pp. 126-133.

Dragan, A., and Akhtar-Danesh, N. (2007), "Relation between body mass index and depression: a structural equation modeling approach," *BMC Medical Research Methodology*, 7:17, 1471-2288.

Enders, W. (1995), *Applied Econometric Time Series* (1st ed.), John Wiley & Sons, Inc.

Faith, M. S., Flint, J., Fairburn, C. G., Goodwin, G. M., Allison, D. B. (2001), "Gender Differences in the Relationship between Personality Dimensions and Relative Body Weight," *Obesity Research*, 9, 10, 647-650.

Faith, M. S., Matz, P. E., and Jorge, M, A. (2002), "Obesity—depression associations in the population," *Journal of Psychosomatic Research*, 53, 935-942.

Flury, B. (1988), *Common Principal Components and Related Multivariate Models*, New York: John Wiley.

Friedman, G. D., Cutter, G. R., Donahue, R. P., Hughes, G. H., Hully, S. B., Jacobs, D. R. J., Liu, K., and Savage, P. J. (1988), "CARDIA: Study design, recruitment, and some characteristics of the examined subjects," *Journal of Clinical Epidemiology*, 41.

Friedman, J. H. (2006), "Recent Advances in Predictive (Machine) Learning," Journal of Classification, 23, 175-197.
Friedman, M. A., and Brownell, K. D. (1995), "Psychological correlates of obesity: moving to the next research generation," *Psychological Bulletin*, 117, 3-20.

Goodman, E., and Whitaker, R. C. (2002), "A Prospective Study of the Role of Depression in the Development and Persistence of Adolescent Obesity," *Pediatrics*, 109, 3, 497-504.

Istavan, J., Zavelta, K., and Weidner, G. (1992), "Body weight and psychological distress in NHANES 1," *International Journal of Obesity*, 16, 999-1003.

Johnson, R. A., and Wichern, D. W. (2002), *Applied Multivariate Statistical Analysis* (5th ed.), Upper Saddle River, NJ: Prentice Hall.

Kakizawa, Y., Shumway, R. H., and Taniguchi, M. (1998), "Discrimination and Clustering for Multivariate Time Series," *Journal of the American Statistical Association*, 93, 328-340.

Kapteyn, A., Neudecker, H., and Wansbeek, T. (1986), "An Approach to N-Mode Components Analysis," *Psychometrika*, 51, 269-275.

Kim, Y., and Adams, B. M. (2009), "Multivariate SPC for Recipe Preservation of Batch Processes," *Quality and Reliability Engineering International*, 26, 3, 267-277.

Korn, F., Jagadish, H. V., and Faloutsos, C. (1997), "Efficiently Supporting Ad Hoc Queries in Large Datasets of Time Sequences," in *ACM SIGMOD International Conference on Management of Data*, Tucson, AZ: ACM, Inc., pp. 289-300.

Kosanovich, K. A., Piovoso, M. J., Dahl, K. S., MacGregor, J. F., and Nomikos, P. (1994), "Multi-Way PCA Applied to an Industrial Batch Process," in *American Control Conference*, Baltimore, MD, pp. 1294-1298.

Kosmelj, K., and Batagelj, V. (1990), "Cross-Sectional Approach for Clustering Time Varying Data," *Journal of Classification*, 7, 99-109.

Lee, J., Yoo, C., and Lee, I. (2004), "Fault detection of batch processes using multiway kernel principal component analysis," *Computers and Chemical Engineering*, 28, 1837-1847.

Liao, T. W. (2005), "Clustering of Time Series Data-a Survey," *Pattern Recognition*, 38, 1857-1874.

Liu, Z., Chen, D., and Bensmail, H. (2005), "Gene Expression Data Classification With Kernel Principal Component Analysis," *Journal of Biomedicine and Biotechnology*, 2, 155-159.

Maharaj, E. A. (2000), "Clusters of Time Series," *Journal of Classification*, 17, 297-314.

Muller, K.-R., Mika, S., Ratsch, C., Tsuda, K., and Scholkpf, B. (2001), "An Introduction to Kernel-Based Learning Algorithms," *IEEE Transactions on Neural Networks*, 12, 181-201.

Negi, T., and Bansal, V. (2005), "Time Series:Similarity Search and Its Applications," *Proceedings - International Conference on Systemics, Cybernetics and Informatics*, 528-533.

Oates, T., Firoiu, L., and Paul, R. C. (1999), "Clustering Time Series with Hidden Markov Models and Dynamic Time Warping," *Proceedings of the IJCAI-99 Workshop on Neural, Symbolic, and Reinforcement Learning Methods for Sequence Learning*.

Onyike, C. U., Crum, R. M., Lee, H. B., Lyketsos, C. G., and Eaton, W. W. (2003). "Is Obesity Associated with Major Depression? Results from the Third National Health and Nutrition Examination Survey," *American Journal of Epidemiology*, 158, 12, 1139-1147.

Palinkas, L. A., Wingard, D. L., and Barrette-Connor E. (1996), "Depressive symptoms in overweight and obese older adults: a test of the 'jolly fat' hypothesis," *Journal of Psychosomatic Research*, 40, 59-66.

Ramoni, M., Sebastiani, P., and Cohen, P. R. (2000), "Multivariate Clustering by Dynamics," *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*.

Ramoni, M., Sebastiani, P., and Cohen, P. R. (2001), "Bayesian Clustering by Dynamics," *Machine Learning*, 1-31.

Roberts, R. E., Kaplan G. A., Shema, S. J., and Strawbridge, W. J. (2000), "Are the Obese at Greater Risk for Depression," *American Journal of Epidemiology*, 152, 2, 163-170.

Roberts, R. E., Deleger, S., Strawbridge, W. J., and Kaplan, G. A. (2003), "Prospective association between obesity and depression: evidence from the Alameda County Study," *International Journal of Obesity*, 27, 514-521.

Sholkopf, B., and Smola, A. J. (2002), *Learning with Kernels -Support Vector Machines, Regularization, Optimization, and Beyond*, Cambridge, MA: The MIT Press.

Shumway, R. H. (1988), *Applied Statistical Time Series Analysis*, Prentice Hall.

Shumway, R. H. (2003), "Time-Frequency Clustering and Discriminant Analysis," *Statistics & Probability Letters*, 63, 307-314.

Simon, G. E., Korff, M. Von., Saunders, K., Miglioretti D. L., Crane, P. K., Belle, G. Van., and Kessler, R. C. (2009), "Association Between Obesity and Psychiatric Disorders in the US Adult Population," *Archives of General Psychiatry*, 63, 824-830.

Stunkard, A. J., Faith, M. S., and Allison, K. C. (2003), "Depression and Obesity," *Society of Biological Psychiatry*, 54, 330-337.
Toshniwal, D., and Joshi, R. C. (2005), "Finding Similarity in Time Series Data by Method of Time Weighted Moments," *Australasian Database Conference (ADC 2005)*.

Wold, S., Esbensen, K., and Geladi, P. (1987), "Principal Component Analysis," *Chemometrics and Intelligent Laboratory Systems*, 2, 37-52.

Wold, S., Geladi, P., Esbensen, K., and Ohman, J. (1987), "Multi-Way Principal Components and Pls-Analysis," *Journal of Chemometrics*, 1, 41-56.