이학석사 학위논문

A Review on Clustering Methods for Functional Data

함수열 자료의 군집방법에 대한 연구

2015년 2월

서울대학교 대학원

통계학과

유 혜 선

A Review on Clustering Methods for Functional Data

지도교수 박 병 욱

이 논문을 이학석사 학위논문으로 제출함

2014년 10월

서울대학교 대학원

통계학과

유 혜 선

유혜선의 이학석사 학위논문을 인준함

2014년 12월

위 원 장	오 희 석	(인)
부위원장	박 병 욱	(인)
위 원	박 태 성	(인)

A Review on Clustering Methods for Functional Data

by

Hyesun Yoo

A Dissertation
submitted in fulfillment of the requirement
for the degree of
Master of Science
in
Statistics

The Department of Statistics
College of Natural Sciences
Seoul National University
February, 2015

Abstract

Hyesun Yoo
The Department of Statistics
The Graduate School
Seoul National University

Many studies have been done for clustering functional data as considerable functional data are obtained recently. We reviewed overall approaches for clustering functional data proposed so far. Those approaches consist of a non-parametric approach which uses dissimilarity between curves as dissimilarity measure, a filtering and clustering technique which is simple and intuitive and a model-based clustering method which assumes a probability distribution of finite dimensional coefficients estimated from data. Model-based methods are reviewed in detail, particularly. Also, we provided an application to energy data using model-based models for functional data to illustrate model-based methods with specific basis.

Keywords : Clustering, Functional data, B-splines, Clustering functional data, Energy usage pattern.

Student Number : 2013-20220

Contents

Chapter 1

Introduction

Clustering analysis searches for homogeneous groups or segmentations of observations. Clustering analysis is often used for explanatory data analysis to identify distinct groups and makes it convenient for the users to interpret and plan intensified research. The hierarchical clustering algorithm and the K-means clustering algorithm have been widely used in many applications.

As considerable amounts of functional-type data are collected recently because of technological advance, demand for clustering such data type is also rising. However, it is often difficult to cluster functional data due to the infinite dimensional space that data belong to. Various approaches have been proposed along the years and people in diverse scientific fields have been searching for a better way with respect to their studies.

Clustering analysis is a type of unsupervised learning which is to infer the properties of X directly without response. One of major obstacles in developing clustering method is that evaluation of such method is intrinsically difficult.

Besides, suitable clustering method should takes into account the context of problems, which make it more ambiguous and complicated. Discussion about those difficulties are well explain in Luxburg, Williamson and Guyon(2012).

An obvious and simple way when we encounter clustering functional data is to consider data as multivariate data and directly conduct clustering algorithm to them. However, this approach may not be informative or applicable to functional data. Functional data could be obtained as densely collected multivariate data type or irregular design data type with discordant recording times or index sets. In addition, measurement error might not be controlled if we do not treat them as functional data. In this case, we have low signal-to-noise ratio which means considerable noise level in the function and it will prevent us from getting stable estimates of a curve.

The purpose of this paper is to review methods of clustering functional data up to date. It is organized as follows. Following introduction section gives necessary backgrounds for both clustering and functional data. Chapter 2 reviews techniques of clustering functional data. Chapter 3 contains how to choose the number of clusters. In chapter 4, an application of the functional clustering methods to energy data is presented. We discuss and summarize our results in chapter 5.

1.1 Basis Expansion

Let Y be a functional random variable from an infinite dimensional space. We assume that we have a set of collected observation $Y_1, \cdots, Y_n$ of Y. Let g_i(t) be the true value for the ith curve at time t and we assume presence of

observational error. Then, we have

$$Y_i(t) \;=\; g_i(t) + \epsilon(t), \qquad i = 1, \;\cdots, n, \tag{1.1}$$

where n is the number of individuals. If ϵ(t) doesn't vary across time, then we can consider it as ϵ for convenience.

The main difficulty that functions are in infinite dimensional space can be handled by using basis expansion method. A basis function is a set of known functions ϕ_k's that are mathematically independent of each other and have the property that we can approximate arbitrarily well any function by taking a weighted sum of linear combination of a sufficiently large number K of these functions. In effect, basis expansion methods approximate a function into a finite-dimensional framework of vectors. Let us consider a basis Φ=$\{\phi_1, \cdots, \phi_K\}$ generating some space and assume that Y admits the the basis expansion

$$g_i(t) = \sum_{k=1}^{K} c_{ik}\phi_l(t). \tag{1.2}$$

Choosing K and basis functions is crucial according to the characteristics of data. The smaller K is, the better the basis functions reflect characteristics of data. Roughness penalty is usually used when we choose K. As for basis functions selection, most functional data analyses in these days involve either a Fourier basis for periodic data, or a B-spline basis for non-periodic data. There is no universally good basis. For instance, wavelet bases are frequently used in energy-related data which show peaks in their daily pattern.(Chaouch (2014)). More contents can be found in Ramsay and Silverman(2005).

1.2 Functional Principal Component Analysis

From the set of functional data, one can be interested in features characterizing typical functions in a parsimonious way. If we use fixed number of basis functions, the eigenfunction basis explains more variation than any other basis expansions. In this respect, functional principal components analysis (FPCA) is an important technique to consider when we cluster functional data. Let X be a L_2 continuous stochastic process,

$$\mu(t) = E(X(t)) \tag{1.3}$$

$$G(s,t) = Cov(X(s), X(t)) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s) \phi_k(t) \tag{1.4}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ are the eigenvalues and ϕ_1, ϕ_2, $\cdots$ are the orthonormal eigenfunctions of the linear Hilbert-Schmidt operator.

By the Karhunen-Loève theorem, one can express the centered process as

$$X(t) - \mu(t) = \sum_{k=1}^{\infty} \xi_k \phi_k(t), \ \textit{where } \xi_k = \int (X(t) - \mu(t))\, \phi_k(t) dt \tag{1.5}$$

Detailed properties and contents could be found in Rogers and Williams(1994). We will not consider FPCA seriously here and regard it as one way of basis expansion methods. It is because FPCA is determined by the data used and our real data application in chapter 4 should not be dependent on data a lot. Therefore, our focus is mainly on basis expansions.

1.3 K-means Algorithm

In clustering method, the goal is to assign close points to the same cluster, a natural loss function would be

$$W(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')=k} d(x_i, x_{i'}). \tag{1.6}$$

Here, we only introduce K-means algorithm which is one of the most popular iterative descent clustering methods. If all variables are of the quantitative type and squared Euclidean distance is chosen as the dissimilarity measure, loss function above can be written as

$$W(C) = \frac{1}{2}\sum_{k=1} K \sum_{C(i)=k}\sum_{C(i')=k} \|x_i - x_{i'}\|^2 \tag{1.7}$$

$$= \sum_{k=1} K N_k \sum_{C(i)=k} \|x_i - \bar{x}_k\|^2, \tag{1.8}$$

where $\bar{x}_k = (\bar{x}_{1k}, \cdots, \bar{x}_{pk})$ is the mean vector associated with the kth cluster, and $N_k = \sum_{i=1}^{N} I(C(i) = k)$. Therefore, this criterion is minimized if we assign the N observations to the K clusters in a way that the average dissimilarity in the cluster is minimized. More contents can be found in Hastie, Tibshirani and Friedman(2009).

Chapter 2

Clustering Functional Data Approaches

A simple way is to regard functional data as high dimensional multivariate data and to apply clustering method such as hierarchical clustering or K-means clustering. This approach which is named as raw-data clustering is conventional multivariate problem and does not take advantage of functional form of the data. In this situation, a main task is to deal with the curse of dimensionality. Consequently, approaches based on dimension reduction, regularization and parsimonious models have been studied. As for details, complete review can be found in Bouveyron and Brunet-Saumard (2014).

In this chapter, three categories are considered. The first category uses dissimilarity between curves as dissimilarity measure. The second category is the most simple and intuitive approach. Finally, the third category is model-based clustering techniques which assume a probability distribution of finite dimensional coefficients estimated from data.

2.1 Nonparametric Approach

Nonparametric approach for functional data clustering is usually a method that applies clustering techniques with specific dissimilarities of distances to smoothed curves. The term, Nonparametric Approach, is referred from Ferraty and Vieu (2006). When people use nonparametric approach, they consider semi-metric based on different kinds of derivative,

$$d_l^{deriv}(x_i, x_{i'}) = \left(\int (x_i^{(l)}(t) - x_{i'}^{(l)}(t))^2 dt \right)^{1/2} \tag{2.1}$$

where $x_i^{(l)}(\mathrm{t})$ is lth derivative of x_i. In usual case, original curves, up to second derivatives of curves or linear combinations of both are used for proximity measure.

However, nonparametric approaches could be reduced to other approaches if we use raw discrete observations or simple proximity measure. For example, if discrete observations of curves are used to measure d_0, it become raw-data clustering methods. In addition, if an approximation of the curves into a finite basis is used to measure d_0, it become filtering and clustering approaches (chapter 2.2) with the same basis.

Due to Parseval's Identity, the L_2 distance between $g_i(\mathrm{t})$ and $g_{i'}(\mathrm{t})$ can be written using notation in (2.2) as

$$d_0(g_i, g_{i'}) = \|g_i - g_{i'}\| = \left(\sum_{k=1}^{\infty} (c_{ik} - c_{jk})^2 \right)^{1/2}. \tag{2.2}$$

Also, Some methods are proposed using new geometric or heuristics criteria to cluster functional data. It is, however, focused on characteristics of spe-

cific data types and is not applicable to general functional data. More basic properties can be found in Peng and Müller (2008).

2.2 Filtering and Clustering Approach

Filtering and clustering approach consists of two steps. As with filtering step, original infinite-dimension functional data are converted into finite-dimension data using basis functions. One often use basis expansion and functional principal component analysis in this step. Next, regarding their coefficients as summary of data, K-means algorithm can be performed in clustering step. Other clustering techniques such as Self-Organised Map and hierarchical clustering can be used as well even though K-means clustering is the most popular.

Filtering and clustering approach is the beginning step for clustering functional data from the functional perspective. Abraham, Cornillon, Matzner-Lber and Molinari(2003) propose this approach first using B-splines and K-means algorithm. Later, Serban and Wasserman(2005) suggest CATS which is adaptable for lasge-scale data by using Fourier transformation and K-means algorithm. These approaches lead to more advanced approaches in next subchapter. However, these approaches have its own advantage that computation is rather easy comparing to those in model-based methods involving EM algorithm. Classification table of papers as type of basis functions and clustering method are presented in Jacques and Preda (2013).

2.3 Model-based Approach

The earliest model-based clustering approach is proposed by James and Sugar (2003). They used a mixed effects spline model to cluster curves. We will follow details of his paper since it shows a general and applicable framework to other circumstances.

Let $x_1, \cdots, x_n$ be observations from mixture distribution with G components and $f_k(x|\theta_k)$ be the density of kth cluster. Also, let $z_i = (z_{i1}, \cdots, z_{iG})$ be the cluster membership vector for the ith observation where z_{ik}=1 if ith observation belongs to kth cluster and 0 otherwise. In the "classification likelihood" approach, take z_i's as parameters, and maximizing the likelihood

$$L_C(\theta_1, \cdots, \theta_G; z_1, \cdots, z_n | x_1, \cdots, x_n) = \prod_{i=1}^{n} f_{z_i}(x_i|\theta_{z_i}). \tag{2.3}$$

Alternatively, the cluster memberships can be treated as missing data where z_i is multinomial with probability ($\pi_1, \cdots, \pi_G$). In this "mixture likelihood" approach, we estimate parameters by maximizing the likelihood

$$L_M(\theta_1, \cdots, \theta_G; z_1, \cdots, z_n | x_1, \cdots, x_n) = \prod_{i=1}^{n} \sum_{k=1}^{G} \pi_k f_k(x_i|\theta_k). \tag{2.4}$$

Using notation (1.1) and (1.2), we let

$$g_i(t) = \boldsymbol{s}(t)^T \boldsymbol{\eta}_i, \tag{2.5}$$

where $\boldsymbol{s}$(t) is a p-dimensional spline basis vector and $\boldsymbol{\eta}_i$ is a vector of spline coefficients. Here, the $\boldsymbol{\eta}_i$'s are modeled using a Gaussian distribution,

$$\boldsymbol{\eta}_i = \boldsymbol{\mu}_{z_i} + \boldsymbol{\gamma}_i, \quad \boldsymbol{\gamma} \sim N(\mathbf{0}, \Gamma), \tag{2.6}$$

where z_i is the unknown cluster membership. A further parametrization of the cluster could be conducted to produce low-dimensional representations of the curves. Note that $\boldsymbol{\mu}_k$ can be presented as

$$\boldsymbol{\mu}_k = \boldsymbol{\lambda_0} + \boldsymbol{\Lambda\alpha_k}, \tag{2.7}$$

where $\boldsymbol{\lambda_0}$ and $\boldsymbol{\alpha_k}$ are p- and h-dimensional vectors and $\boldsymbol{\Lambda}$ is a p $\times$ h matrix with h $\leq$ min(p,G-1).

With above formulations, the functional clustering model (FCM) can be written as

$$\boldsymbol{Y}_i = S_i(\boldsymbol{\lambda_0} + \boldsymbol{\Lambda\alpha_k} + \boldsymbol{\gamma}_i) + \boldsymbol{\epsilon}_i, \quad i = 1, \cdots, n, \tag{2.8}$$

$$\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, R), \quad \boldsymbol{\gamma}_i \sim N(\mathbf{0}, \Gamma), \tag{2.9}$$

where S_i=$(\boldsymbol{s}(t_{i1}), \cdots, \boldsymbol{s}(t_{in_i}))^T$ is the spline basis matrix for the ith curve. For simplicity, we will denote the functional clustering model by FCM. EM algorithm is used to fit the model and estimate $\boldsymbol{\lambda_0}$, $\boldsymbol{\Lambda}$, $\boldsymbol{\alpha_k}$, Γ ,and σ^2. This process involves iterative procedure. Details can be found in James and Sugar (2003).

The use of spline basis is convenient, but sometimes is not appropriate for data that show peaks. Thus, Giacofci, Lambert-Lacroix, Marot and Picard (2013) recently develop a model-based method using wavelet basis. They made their procedure more adaptable for high dimensional setting by including dimensionality reduction step which is useful for screening out less useful coefficients.

Model-based approaches based on principal components techniques also have been studied. Since dimension reduction is a crucial component in clustering functional data, FPCA serves as nice basis given data. A broad outline

of their methods is assuming Gaussian distribution of the principal components and defining model-based clustering techniques as the following mixture model,

$$L_M(\theta; z_1, \cdots, z_n | x_1, \cdots, x_n) = \sum_{k=1}^{G} \prod_{i=1}^{q_k} \pi_k f_{C_j}(c_{jk}(x) | \lambda_{jk}). \tag{2.10}$$

where $\theta = (\pi_k,\ \lambda_{1k},\ \cdots,\ \lambda_{q_k k}\ )_{1 \le k \le K}$ are the parameters and q_k is the order of truncation of the the Karhunen-Loève expansion. Details of these methods can be found in Chiou and Li (2007).

Chapter 3

Choice of the number of clusters

In cluster analysis, determining the number of clusters is the one of the difficult problem. Banfield and Raftery (1993) used a derived approximation to twice the log Bayes factors, which are difficult to calculate. Therefore, Bayes information criterion(BIC) which is an approximation of twice the log Bayes factors is a popular choice.

$$-2log(p(\boldsymbol{x}|\theta)) + constant \tag{3.1}$$

$$\approx -2l(\boldsymbol{x}, \hat{\theta}) + mlog(n) \equiv BIC \tag{3.2}$$

where $p(\boldsymbol{x}|\theta)$ is the likelihood of the data, $l(\boldsymbol{x},\ \hat{\theta})$ is the maximized mixture log likelihood for the model and m is the number of independent parameter to be estimated in the model. Also, n is the number of curves. The smaller the value of the BIC, the stronger the evidence for the model.

Sugar and James (2003) also give an alternative approach based on the "distortion function"

$$d_K = \frac{1}{p} min_{c_1,\cdots,c_K} E(\boldsymbol{\eta}_i - \boldsymbol{c}_{z_i})\Gamma^{-1}(\boldsymbol{\eta}_i - \boldsymbol{c}_{z_i}), \tag{3.3}$$

where $\boldsymbol{\eta}_i$'s are the spline coefficients in the functional clustering model. The distortion, d_K, is the average Mahalanobis distance between each $\boldsymbol{\eta}_i$ and its closest cluster center, c_{z_i}. Then the largest jump $\mathrm{d}_K^{-1} - d_{K-1}^{-1}$ will determine the number of clusters. There are also other methods such as gap statistics and ICL. However, We will only consider BIC in chaper 4.

Chapter 4

Application to real data

4.1 Data Description

Energy analysis on aggregate data in building level or distribution panels has traditionally been the most primary interest for building energy management. However, nowadays, with the advent of Smart Grid technologies, it is possible to collect the data at a finer level and this enables us to analyze energy usage in hierarchical levels from building-level to equipment-level energy consumption. By learning energy consumption patterns of various energy equipment connected to sensors, we can detect the problems like energy leak as well as we can develop an efficient management plan even for energy equipment in buildings. In addition, clustering similar household-level energy consumption is an issue in energy industry in that they can provide adequate energy when it is needed. Also, if we know energy consumption patterns, we can infer individual life pattern accordingly.

The data we used for this study was collected from 18 panel boards out

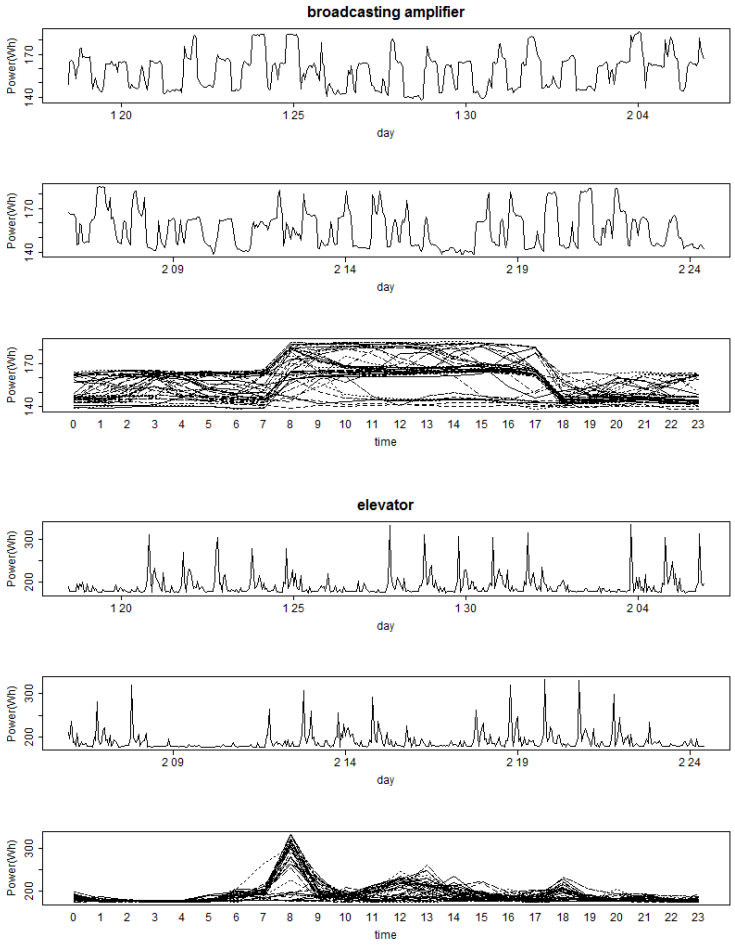

Figure 4.1: broadcasting amplifier(above three plots) and elevators(below three plots); plots of each channel consist of 2 time series plots and an interaction plot.

of 256 power distribution panel boards deployed in an apartment model house building in Seoul, South Korea, over 37 days from January 19, 2013 to February 24, 2013. Thirty seven days consist of 13 weekend days and 24 week days. Each panel board has a sensor board that measures the electricity usage from 24 metering channels through sensors attached to subsidiary power circuits and a process board that is in charge of sending the collected sensor measurements in every second through network to central server. Data were provided by Bell Las Seoul and Gachon University.

Several channels show similarities in their usage patterns. Also, some channels have several typical Intra-day patterns. We applied model-based clustering method to each channel. In this paper, We presented only two selected channels which seem to have different-type curves and are not separated apparently. One of these is used to provide energy for broadcasting amplifier;the other is for elevators.

4.2 Results

We used a model-based method for functional data using two kinds of basis. One is B-spline basis; the other is wavelet basis which is suitable for energy data of which detecting energy peak is important characteristics. R-codes for B-spline are available on the James's website; R package "curvclust" for wavelet on R Archive. Choosing the number of basis is also a difficult issue. Therefore, we specified that the dimension of B-spline basis is ten for simplicity. FCM with basis less then 10 performs well, whereas mean curves with this basis are not reflecting daily patterns sufficiently. In curvclust, everything is automatically chosen according to BIC. We used default values since our main interest is

clustering of functional data. Therefore, performance of two methods could not be compared precisely. However, it is still meaningful to have two results from different methods since there is no exact evaluation standard in real data problems.

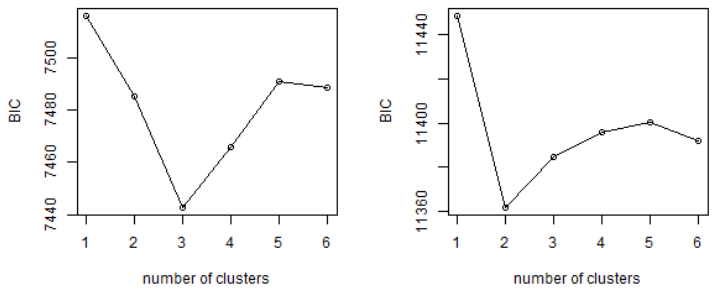

Figure 4.2: BIC obtained from FCM with B-spline for broadcasting amplifier(left) and elevator(right)

Broadcasting Amplifier

The number of clusters is determined by using BIC. Results of the BIC are presented in Figure 4.2. It seems that three is the most reasonable choice according to BIC. In curvclust package, wavelet basis is automatically selected. Therefore, it is difficult to get comparable BIC in this package. Therefore, BIC chosen by FCM with B-splines is used for both cases. Figure 4.3 shows clustering results. Curves in same clusters are drawn together in dotted lines and each mean curve is drawn in bold line. It seems that homogeneous curves are clustered together in both methods. What is worth of notice is that mean curves using FCM with wavelet capture sudden drop or sudden increase very well unlike FCM with B-spline. In addition, Table 4.1 suggests that there is

significant difference in energy use pattern between weekend and week day, which is normally expected in energy consumption data. There would be different energy consumption patterns between weekend and week. Curves in cluster 1 are exactly same and difference between cluster 2 and 3 may attributed to basis types. Figure 4.4 is a clustering result which uses a method for visualizing time series. Color gets darker when its value gets higher. If there is significant less variation, at least we can say clustering made data ordered which is usually main purpose of explanatory data analysis (EPA).

	FCM(B-spline)			FCM(wavelet)		
Cluster	cluster 1	cluster 2	cluster 3	cluster 1	cluster 2	cluster 3
Week day	0	15	9	0	10	14
Weekend	7	6	0	7	1	5

Table 4.1: Type of days within each cluster of broadcasting amplifier data

Elevator

Energy use pattern in elevator has high peak during the morning rush hours around eight a.m. Also, there are small peaks during mealtime around 1 a.m and 6 p.m. Results of the BIC are presented in Figure 4.2 and we choose two clusters according to BIC. Figure 4.5 shows clustering results. FCM with wavelet explained mean curves of energy use in peak time more appropriately than FCM with B-spline in Figure 4.5. However, we can detect abnormal behavior of curves during 1 a.m to 7 a.m. In this channel, both methods give same results. In table 4.2, they perfectly discern week day and weekend pattern. It is because this channel has distinctive weekly pattern.

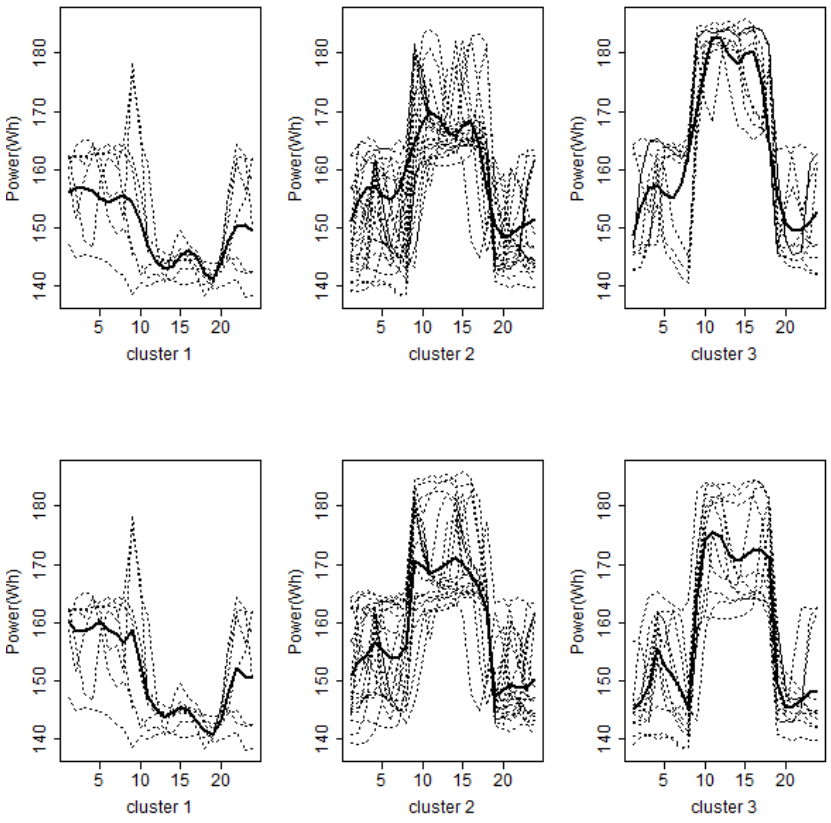

Figure 4.3: Clusters obtained for broadcasting amplifier using FCM with B-spline(top) and FCM with wavelet(bottom)

	FCM(B-spline)		FCM(wavelet)	
Cluster	cluster 1	cluster 2	cluster 1	cluster 2
Week day	0	24	0	24
Weekend	13	0	13	0

Table 4.2: Type of days within each cluster of elevator

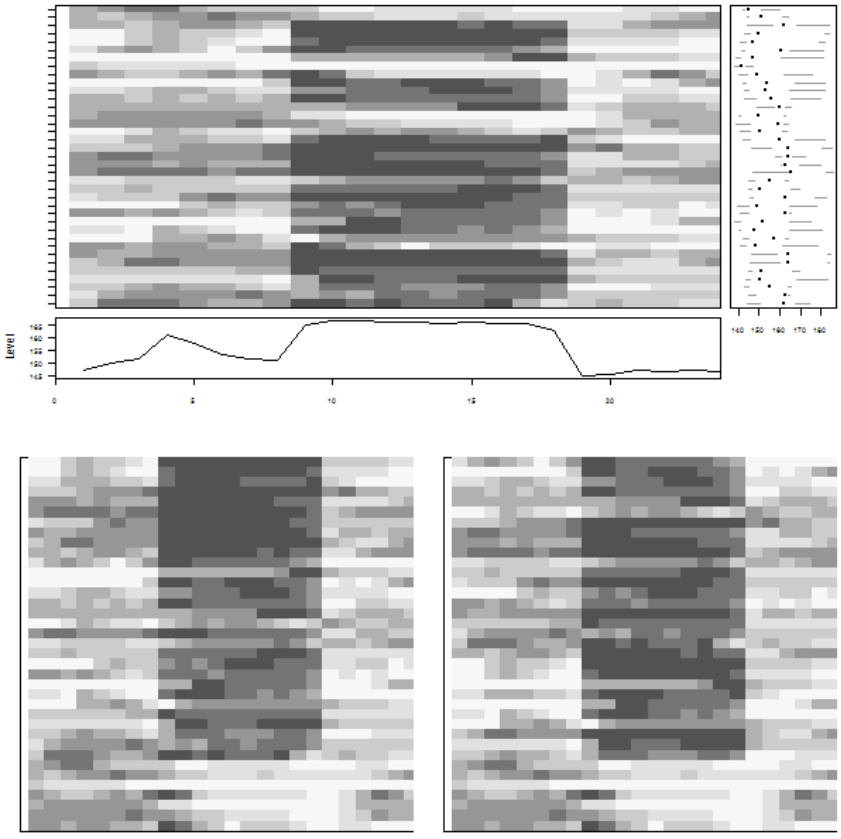

Figure 4.4: Original daily energy data of broadcasting amplifier(above) and clustered energy data of broadcasting amplifier using FCM with B-spline(below, left); using FCM with wavelet(below, right)

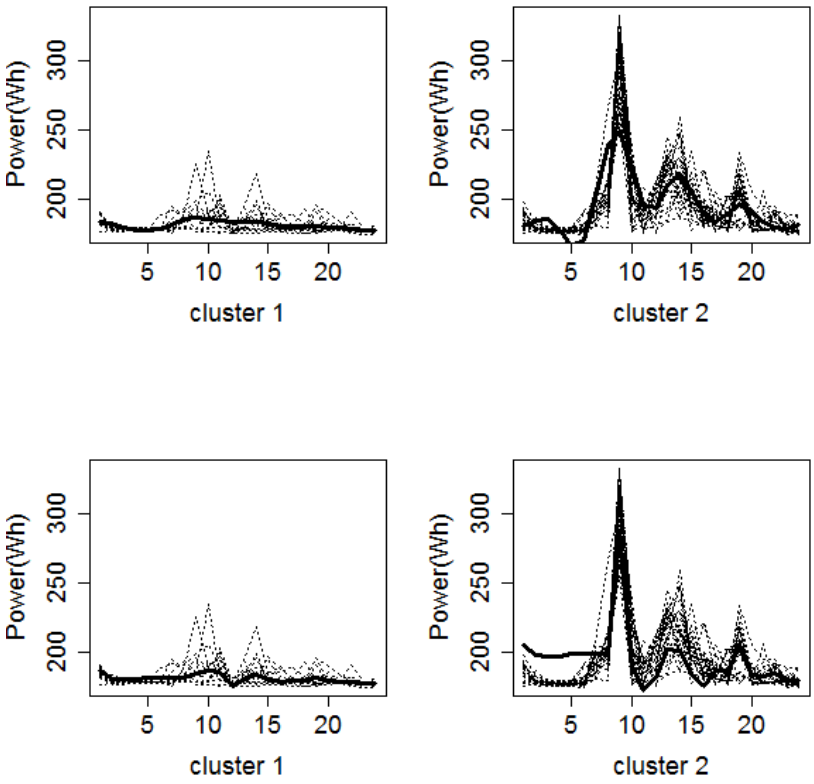

Figure 4.5: Clusters obtained for elevator using FCM with B-spline(top) and FCM with wavelet(bottom)

Chapter 5

Conclusion

In this paper, We review clustering functional methods according to three main frameworks. The first one is a nonparametric approach which uses various kinds of proximity measures. The second one is filtering and clustering approach which has been studied actively before the third one came into spotlight. The third one is a model-based approach which does filtering and clustering simultaneously. Each category has its own strengths and weaknesses. Therefore, we should choose appropriate methods considering goals of research and characteristics of data. Despite of pros and cons of each method, model-based methods can be suggested for several reasons. They perform filtering and clustering simultaneously; computing diverse proximity measures is difficult in respect of calculation and selection of basis. Clustering functional data based on model-based methods are actively studied recently. We also consider the problem of choosing the number of clusters. We presented an application of model-based methods to energy data for illustration.

We applied FCM to channels' energy consumption data. Even though

both FCM with B-spline and FCM with wavelet well performed for clustering in energy data, mean curves using FCM with wavelet basis depict energy use more properly. Therefore one can use both methods, if clustering is a purpose of research. FCM with B-spline could be more general choice due to flexibility of basis. One is recommended to use wavelet basis, however, if an acquisition of proper mean curves is an major purpose. In case of some bumps or peaks being determining factors, wavelet basis is also suggested. This research on energy data can be extended to clustering individually similar energy sources, which is one of promising topics in energy industry as more and more smart-meters are distributed.(Kwac, Flora and Rajagopal (2014)). Thus, problem of clustering daily patterns within an individual and clustering homogeneous individuals simultaneously should be further studied in the near future.

References

Abraham, C., Cornillon, P.A., Matzner-Lber, E., and Molinari, N. (2003), "Unsupervised Curve Clustering Using B-Splines," *The Scandinavian Journal of Statistics* **30** ,581–595.

Banfield, J.D., and Raftery, A.E. (1993), "Model-based Gaussian and non-Gaussian Clustering," *Biometrics* **49**, 803–821.

Bouveyron, C. and Brunet-Saumard, C., (2014), "Model-Based Clustering of High-Dimensional Data: A Review," *Computational Statistics and Data Analysis*, **71**, 52–78.

Chiou, J.M. and Li, P.L. (2007), "Functional Clustering and Identifying Substructures of Longitudinal Data," *Journal of the Rotal Statistical Society, Series B,* **69**, 679–699.

Ferraty, F. and Vieu, P. (2006), *Nonparametric Functional Data Analysis,* Springer, New York.

Fraley, C. and Raftery, A.E., (1998), "How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis" *The Computer Journal*, **41**, 578–588.

Giacofci, M., Lambert-Lacroix, S., Marot, G. and Picard, F. (2013), "Wavelet-Based Clustering for Mixed-Effects Functional Models in High Dimension, " *Biometrics*, **69**, 31–40.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning,*, Second Edition Springer, New York.

Ieva, F., Paganoni, A.M., Pigoli, D. and Vitelli, V. (2012), "Multivariate Functional Clustering for the Analysis of ECG Curves Morphology," *Journal of the Royal Statistical Society, Series C,* **62**, 2012

Jacques, J. and Preda, C. (2013), "Functional data clustering: a survey," Research Report, informatics mathematics.

James, G.M and Sugar, C.A. (2003), "Clustering for Sparsely Sampled Functional Data," *Journal of American Statistical Association*, **98**, 397–408.

Luxburg, U.V., Williamson, R.C. and Guyon, I. (2012), "Clustering: Science or Art," *JMLR: Workshop and Conference Proceedings*, **27,** 65–79.

Mohamed Chaouch. (2014), "Clustering-Based Improvement of Nonparametric Functional Time Series Forecasting: Application to Intra-Day Household-Level Load Curves," *IEEE TRANSACTIONS ON SMART GRID*, **5**, 411–419.

Peng, J. and Müller, H.G. (2008), "Distance-based Clustering of Sparsely Observed Stochastic Processes, with applications to online auctions," *The Annals of Applied Statistics,***2**, 1056-1077.

Peng, R.D. (2008), "A Method for Visualizing Multivariate Time Series Data," *Journal of Statistical Software*, **25**, Code snippet1.

Ramsay, J.O. and Silverman, B.W. (2005), *Functional Data Analysis,* Second Edition, Second Edition, Springer, New York.

Rogers, L.C.G. and Williams, D.(1994), *Diffusions, Markov processes, and Martingales*, Second Edition, Cambridge University Press, Cambridge.

Serban, N., and Wasserman, L., "CATS: Clustering After Tranformation and Smoothing," *Journal of the American Statistical Association* **100,** 990–999.

국문초록

함수열 자료들의 초기 분석 단계나 전처리 단계 수준에서 대표적인 곡선의 형태나 모양을 얻는 일이 필요해지면서 함수열 자료의 군집방법에 대한 연구가 활발히 진행되고 있다. 이러한 연구는 앞으로 점점 함수열 자료가 많이 발생함에 따라 더욱 더 활발해질 전망이다. 이 논문에서는 함수열 자료의 군집방법에 대한 전반적인 내용에 대하여 각 특징별로 분류하고 검토하였다. 특히 가장 전반적으로 적용 가능한 모형 기반 군집방법을 자세히 살펴보았다. 또한 모형 기반 군집방법을 빌딩 내의 각 특정한 장치에서 얻어진 시간별 에너지 데이터에 적용한 결과를 제시하였다.

주요어 : 함수열 자료, 군집방법, 함수열 자료의 군집방법, B-spline

학　번 : 2013-20220