Linkage Based Dirichlet Processes

Yuhyun Song

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Scotland C. Leman, Chair
Leanna L. House
Inyoung Kim
George R. Terrell

January 27, 2017
Blacksburg, Virginia

Keywords: concentration parameter, Dirichlet processes, nested Dirichlet processes

Linkage Based Dirichlet Processes

Yuhyun Song

(Abstract)

We live in the era of *Big Data* with significantly richer computational resources than the last two decades. The concurrence of computation resources and a large volume of data has boosted researchers' desire for developing feasible Markov Chain Monte Carlo (MCMC) algorithms for large parameter spaces. Dirichlet Process Mixture Models (DPMMs) have become a Bayesian mainstay for modeling heterogeneous structures, namely clusters, especially when the quantity of clusters is not known with the established MCMC methods. As opposed to many ad-hoc clustering methods, using Dirichlet Processes (DPs) in models provide a flexible and probabilistic approach for automatically estimating both cluster structure and quantity. While DPs are not fully parameterized, they depend on both a base measure and a concentration parameter that can heavily impact inferences.

Determining the concentration parameter is critical and essential, since it adjusts the a-priori cluster expectation, but typical approaches for specifying this parameter are rather cavalier. In this work, we propose a new method for automatically and adaptively determining this parameter, which directly calibrates distances between clusters through an explicit link function within the DP. Furthermore, we extend our method to mixture models with Nested Dirichlet Processes (NDPs) that cluster the multilevel data and depend on the specification of a vector of concentration parameters. In this work, we detail how to incorporate our method in Markov chain Monte Carlo algorithms, and illustrate our findings through a series of comparative simulation studies and applications.

Linkage Based Dirichlet Processes

Yuhyun Song

(General Audience Abstract)

We live in the era of *Big Data* with significantly richer computational resources than the last two decades. The concurrence of computational resources and a large volume of data has boosted researcher's desire to develop the efficient Markov Chain Monte Carlo (MCMC) algorithms for models such as a Dirichlet process mixture model. The Dirichlet process mixture model has become more popular for clustering analyses because it provides a flexible and generative model for automatically defining both cluster structure and quantity. However, a clustering solution inferred by the Dirichlet process mixture model is impacted by the hyperparameters called a base measure and a concentration parameter.

Determining the concentration parameter is critical and essential, since it adjusts the a-priori cluster expectation, but typical approaches for specifying this parameter are rather cavalier. In this work, we propose a new method for automatically and adaptively determining this parameter, which directly calibrates distances between clusters. Furthermore, we extend our method to mixture models with Nested Dirichlet Processes (NDPs) that cluster the multilevel data and depend on the specification of a vector of concentration parameters. In this work, we have simulation studies to show the performance of the developed methods and applications such as modeling the timeline for building construction data and clustering the U.S median household income data.

This work has contributions: 1) the developed methods in this work are straightforward to incorporate with any type of Monte Carlo Markov Chain algorithms, 2) methods calibrate with the probability distance between clusters and maximize the information based on the observations in defined clusters when estimating the concentration parameter, and 3) the methods can be extended to any type of the extension of Dirichlet processes, for instance, hierarchical Dirichlet processes or dependent Dirichlet processes.

Dedication

To my precious family.

Acknowledgments

I would like to thank a number of people who supported and guided me through my PhD. First of all, I would like to express my sincere gratitude to my advisor Scotland Leman for his persistent guidance, patience, and support. Thanks to his support and advice, I have been able to overcome many difficult times. I have been lucky to have him as my advisor.

Besides my advisor, I would like to express the appreciation to my doctoral committee members, Dr. House, Dr. Terrell, and Dr. Kim. Thanks to their constructive suggestions and encouragement, this dissertation would have been achievable. In addition, a special thanks to Dr. Kim for taking care of me like a family. I appreciate your time on listening to my concerns and giving me productive advice.

My dear friends, Lucas Roberts, Xinran Hu, Andy Hoegh, Marcos Carzolio, Ian Crandell, J.T. Fry, Matt Slifko, and Nathan Wycoff, I have enjoyed our Friday meeting because of all of you. Also, thanks to all of my friends in the department of statistics and other department, my life as a graduate student is enjoyable. Thank you!

Finally, I would like to thank my family for your faith in me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Grouping objects is an inherent aspect of human activity. People distinguish objects based on their features and characteristics, and then segment them into groups. In statistics and computer sciences, the task of grouping is referenced to as a clustering analysis. As clustering analyses have gained currency in several areas, various clustering methods have been developed and widely used in many areas for segmenting data objects based on similarities (or dissimilarities). These disciplines include, but are not limited to biology, psychology, social science, and medicine. For instance, in the area of marketing, the goal of clustering is to find apt target consumer groups so that a company can execute an effective marketing campaign (Linoff and Berry, 2011). As an example, a clustering method assigns genes into natural groups, providing similar functions (Ashburner et al., 2000). Also, with the growing number of social networking platforms like Facebook and Twitter, many researchers are interested in grouping users and identifying the homogeneous (or heterogeneous) behaviors of users through employing various clustering methods (Wakita and Tsurumi, 2007). Clustering analyses have also been used for crime analyses in order to determine locations with high incidence of crime (Chen et al., 2004; Murray et al., 2001). With the surge of interest in *Big Data*, researchers use clustering algorithms for text mining which involve classifying huge quantifies of documents from various sources, such as Twitter, Facebook, and news (Dhillon

Figure 1.1: Visualization of Grouping objects. Image is from https://blogs.stthomas.edu/hphc/2012/05/04/segmenting-the-future-of-health-care.

and Modha, 2001; Steinbach et al., 2000; Blei et al., 2003).Clearly, clustering is becoming the cornerstone of many analyses. However, the uncertainty in the number of clusters has been overlooked in the use of ad-hoc clustering algorithms.

Among various clustering methods, agglomerative hierarchical clustering approaches utilize linkage functions to measure distances between clusters. In general, these methods first determine a pair of closest clusters, and merge them at each stage (Kaufman and Rousseeuw, 2009; Hastie et al., 2009). Agglomerative hierarchical clustering algorithms do not require the number of clusters to be fixed in advance. However, in order to determine the number of clusters, users inspect a resulting dendrogram which graphically lays out clusters and the distances between them. Therefore, agglomerative hierarchical clustering algorithms have a disadvantage, in that a clustering solution in accordance with the number of clusters may be biased depending upon the researcher's subjective view.

The K-means clustering algorithm requires that objects are partitioned into a predetermined number of K clusters, centered around K centroids, by minimizing the within cluster

sum of squares (Hartigan and Wong, 1979). However, in order to use the K-means clustering approach, users are required to select the number of clusters in advance and investigate the selected number of clusters, for instance, by using a scree plot, which is a graphical tool for determining K. Many such distance function modifications, e.g. K-medoids algorithm (Kaufman and Rousseeuw, 1987), have been adapted for added flexibility in suggesting what a cluster might look like. However, both agglomerative hierarchical clustering and K-means clustering are inflexible for individual clusters. Also, once observations are assigned to clusters, clustering results cannot be easily changed. Moreover, an inference about the number of clusters and the quality of clustering cannot be assessed parametrically, because neither method provides any generative models.

A Gaussian Mixture Model (GMM) solves some of the concerns associated with the K-means clustering approach and agglomerative hierarchical clustering algorithms. Similar to K-means clustering, the GMM is often used for clustering purposes by assuming that the observations are drawn from the mixture of Gaussian distributions with mean μ_i and the standard deviation s_i where $i = 1\,, \ldots, K$, such that K is the number of clusters (or mixture components) in the GMM (McLachlan and Basford, 1988; Figueiredo and Jain, 2002; McLachlan and Peel, 2004). Like K-means clustering, when using this approach, the number of mixture components should be predetermined prior to fitting the model. Without prespecified information, it is not easy to choose the number of components or clusters, and it requires further investigations to examine the selected number of clusters. However, unlike the K-means clustering approach and agglomerative hierarchical clustering algorithms, an inference about the number of clusters can be made by model selection methods, such as the Akaike Information Criterion (AIC) (Akaike, 1974), the Bayesian Information Criterion (BIC) (Zhou and Hansen, 2000; Schwarz et al., 1978; Chen and Gopalakrishnan, 1998), the Integrated Completed Likelihood (ICL) (Biernacki et al., 2000), and the Likelihood Ratio Test (LRT) (McLachlan, 1987).

Recently, Dirichlet process priors for mixture modeling have taken center stage as providing both the identity of cluster labels and the number of clusters simultaneously (Antoniak, 1974; MacEachern and Müller, 1998). Unlike ad-hoc clustering methods, the Dirichlet process prior enables us not only to bypass the issue of predetermining the number of clusters, but also to make inferences about clustering results. For example, Dahl (2006) uses a Dirichlet process mixture model as a model-based clustering in order to group the gene expression data. Consequently, clustering results provided by the Dirichlet process mixture model automatically offer inferences about both the number of clusters and model parameters. Also, the use of the Dirichlet process prior is common-place for handling heterogeneity in regression coefficients and determining the number of clusters (Kim et al., 2004). Assuming that each cluster of data has its own mixture model, Teh et al. (2006) suggests building a semi-parametric approach to model groups of data hierarchically. This approach involves constructing a hierarchical Dirichlet process for groups of data and applying their method into text modeling.

A Dirichlet process prior is formed by two hyperparameters: a concentration parameter (or a strength parameter) α, and a base measure G_0. It is common knowledge that clustering results are sensitive to selections of these two hyperparameters (West, 1992; McAuliffe et al., 2006; Rabaoui et al., 2011). Thus, we need to develop dependable approaches for estimating hyperparameters in a Dirichlet process in order to ensure efficient and accurate clustering results. Especially, it is known that unknown concentration parameter α plays an important role because it determines the expected number of clusters in the Dirichlet process. In this work, we propose a new method for estimating the concentration parameter, which has an effect on the expected number of clusters in the DP. Unlike other previous methods for estimating the concentration parameter, our method uses distances between clusters, which implies that we borrow and pool the information from defined clusters.

This work is organized as follows. In Chapter 2, we review the Dirichlet process and the Dirichlet process mixture model. We then introduce our new method, linkage based Dirichlet

processes, in Chapter 3. Then, as a case study, we apply our method to simulated data under different cases and compare with the result from another method proposed by McAuliffe et al. (2006). As an application, we implement a mixture model with a linkage based Dirichlet process prior for modeling the timeline for building construction costs in Chapter 5. In Chapter 6, we review nested Dirichlet processes and propose a linkage based nested Dirichlet process, which extends linkage based Dirichlet process to the nested Dirichlet process. This is followed by a simulation study for a linkage based nested Dirichlet process mixture model in Chapter 7. Finally, in Chapter 8, we conclude and discuss future works.

Chapter 2

Overview of Dirichlet Processes

2.1 Dirichlet Processes

A Dirichlet Process (DP), which is named after *Peter Gustav Lejeune Dirichlet*, is the stochastic process that is used in Bayesian inferences (Blackwell and MacQueen, 1973). The Dirichlet process is often called "a distribution over distributions" because this stochastic process describes the prior knowledge about the distribution of random variables in Bayesian semi-parametric models and each draw from this stochastic process is a realization of probability distribution (Blackwell and MacQueen, 1973; MacEachern and Müller, 1998; McAuliffe et al., 2006; Liu, 1996). The DP has gained popularity in machine learning, computer science, and statistics due to its properties, such as flexibility and clustering effects. Particularly, in linear mixed models, the DP prior is often assigned for the distribution of random effects when they do not follow certain parametric distributions such as a normal distribution (Kleinman and Ibrahim, 1998; Mukhopadhyay and Gelfand, 1997). In unsupervised learning, in a Dirichlet process mixture model, the DP is used as the prior on the number of clusters for clustering analyses (MacEachern and Müller, 1998; Escobar, 1994). Then, the DP allows us to perform clustering analyses without possessing information about the number of clusters due to its potentially infinite nature. In Chapter 2, we briefly review

Dirichlet distributions, Dirichlet processes, Dirichlet process mixture models, and Markov chain Monte Carlo algorithms for implementing Dirichlet process mixture models.

2.1.1 Dirichlet Distributions

A Dirichlet distribution is the probability distribution for K-dimensional random vectors $\boldsymbol{x}$, which elements are non-negative number in $[0, 1]$ and $\sum_{i=1}^{K} x_i$ is at most 1 (Blackwell and MacQueen, 1973; Fabius, 1973). $\forall i : x_i \geq 0$ and $\sum_{i=1}^{K} x_i = 1$, the Dirichlet distribution denoted by $\boldsymbol{x} \sim Dir(\boldsymbol{\alpha})$ has the following probability density distribution on Euclidean space $\mathbf{R}^{K-1}$:

$$f\left(x_1, \ldots, x_{K-1}, x_K \mid \boldsymbol{\alpha}\right) = \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} x_i^{\alpha_i - 1},$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$. Uniform distributions and Beta distributions are the special cases of the Dirichlet distribution. That is, when $\alpha_1 = \alpha_2 = \ldots = \alpha_K = 1$, the Dirichlet distribution becomes the uniform distribution. When $K = 2$, the Dirichlet process yields a Beta distribution, $Beta(\alpha_1, \alpha_2)$, for a scalar $\boldsymbol{x}$.

Under a Bayesian framework, a Dirichlet distribution is the conjugate prior for probability parameter $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots, \pi_K\}$ in the multinomial distribution (Ferguson, 1973; Antoniak, 1974; Minka, 2000; Blei et al., 2003). Given data $\boldsymbol{x}$, the posterior distribution of $\boldsymbol{\pi}$ is also a

Dirichlet distribution:

$$\begin{aligned} f\left(\boldsymbol{\pi} \mid \boldsymbol{x}\right) &\propto \left(\frac{n!}{x_1!x_2!\ldots x_K!}\prod_i^K \pi_i^{x_i}\right)\left(\frac{\Gamma(\alpha_1+\ldots+\alpha_K)}{\prod_i^K \Gamma(\alpha_i)}\prod_i^K \pi_i^{\alpha_i-1}\right) \\ &\propto \prod_i^K \pi_i^{\alpha_i+x_i-1} \\ &\propto Dir(\boldsymbol{\alpha}+x). \end{aligned}$$

2.1.2 Formal Definition of Dirichlet Processes

We view a DP as an infinite-dimensional generalization of the Dirichlet distribution. This suggests that the Dirichlet process is useful for modeling a distribution over distributions because the domain of the Dirichlet distribution can be expressed as a set of K discrete probability distributions (Ferguson, 1973; Balakrishnan, 2001). According to the original definition by Ferguson (1973), consider a measurable space Ω and the finite partition $\boldsymbol{A} = \{A_1, A_2, \ldots, A_K\}$ of Ω. Let G be a random distribution over Ω. Since G is random, $(G(A_1), G(A_2), \ldots, G(A_K))$ is a random vector. We defines $G \sim DP(\alpha, G_0)$, where α and G_0 are a concentration parameter and a base measure, if

$$G(A_1), G(A_2), \ldots, G(A_K) \sim Dir(\alpha G_0(A_1), \alpha G_0(A_2), \ldots, \alpha G_0(A_K)),$$

for every finite partition $\boldsymbol{A} = \{A_1, A_2, \ldots, A_K\}$ of Ω.

A DP is determined by a base measure G_0 and a concentration parameter α, which are the probability measure and the positive real number, respectively. In the DP, for any measurable set $A \subset \Omega$, the base measure G_0 and the concentration parameter α have different

roles as follows:

$$E\left(G(A) \mid G_0, \alpha\right) = G_0(A),$$
$$Var(G(A) \mid G_0, \alpha) = \frac{G_0(A)(1 - G_0(A))}{\alpha + 1}.$$

The base measure (G_0), which is itself a distribution, is an expectation of the Dirichlet process, which implies that the Dirichlet process samples the distributions around the base measure. The concentration parameter (α) has an influence on determining the variance of the DP. If α increases, the DP exhibits the smaller variance, then samples drawn from the DP are more likely to concentrate on its mass around the mean $G_0(A)$ (Teh (2010)).

Now, we explain the posterior distribution of the Dirichlet process. Given by the random distribution $G \sim DP(\alpha, G_0)$, let an exchangeable sequence $\theta = \{\theta_1, \theta_2, \ldots, \theta_N\}$ follow the random distribution G over a measurable space Ω. For the finite measurable partition $\{A_1, A_2, \ldots, A_K\}$ of Ω, let n_k be the number of observed values of $\theta_1, \theta_2, \ldots, \theta_N$ in A_k for $i = 1, 2, \ldots, N$ and $k = 1, 2, \ldots, K$. Due to the conjugacy between a Dirichlet distribution and a multinomial distribution, the posterior distribution of the DP becomes:

$$G(A_1), G(A_2), \ldots, G(A_K) \sim Dir(\alpha G_0(A_1) + n_1, \ldots, \alpha G_0(A_K) + n_K), \tag{2.1}$$

where $\delta(\theta = \theta_i)$ is the Dirac delta centered at θ_i. Equation 2.1 shows that the posterior distribution of the DP is another Dirichlet process with $\alpha^* = \alpha + N$ and $G_0^* = \frac{\alpha G_0 + \sum_{i=1}^{N} \delta(\theta = \theta_i)}{\alpha + N}$ (Görür, 2007). Also, the posterior predictive distribution for θ_{N+1} is described as

$$\theta_{N+1} \,|\theta_1,\, \theta_2, \ldots, \theta_{N-1},\, \theta_N, \sim \frac{1}{\alpha + N}\left(\alpha G_0(\theta_{N+1}) + \sum_{i=1}^{N} \delta(\theta_{N+1} = \theta_i)\right). \tag{2.2}$$

A base measure can be continuous, but the fact of the matter is that the sampled distributions from a DP are discrete probability measures with the probability one (Blackwell and

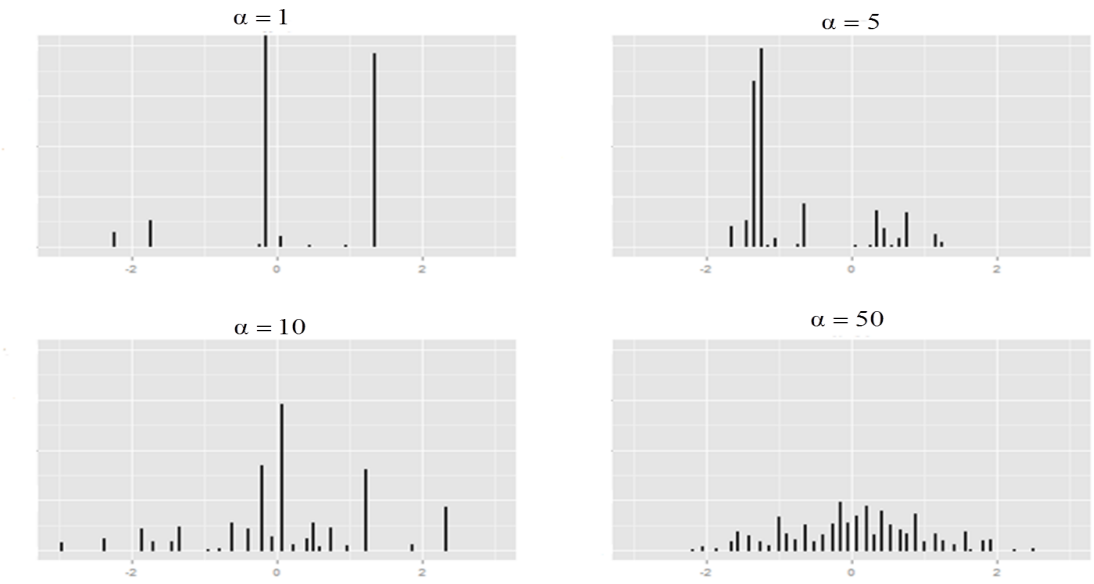


Figure 2.1: 1000 samples drawn from $G \sim DP(\alpha, G_0)$ with 4 different values of the concentration parameter $\alpha = 1, 5, 10, 50$, where $G_0 \sim N(0, 1)$.

MacQueen (1973)). In addition, the conditional distribution in Equation 2.2 implies that a DP naturally exhibits clustering effects (Antoniak, 1974; Aldous, 1985; Teh, 2010). Let $\theta^* = \{\theta_{1^*}, \theta_{2^*}, \ldots, \theta_{K^*}\}$ be the distinct values of $\theta = \{\theta_1, \theta_2, \ldots, \theta_N\}$. Then, $\theta_1, \theta_2, \ldots, \theta_N$ are naturally grouped into K^* clusters.

Figure 2.1 depicts samples drawn from the Dirichlet process. The locations at which peaks along the x-axis occur correspond to the values of samples drawn from the DP. This demonstrates that the samples from the $DP(\alpha, G_0)$ are discrete. Also, the number of distinct values of samples is the same as the number of peaks. It is clear that the number of distinct values of samples increases as α increases. This illustrates that the number of clusters depends on the size of the concentration parameter. These properties of the DP makes this stochastic process more popular in Bayesian modelings, particularly for infinite mixture models (Rasmussen, 1999; Neal, 2000; Medvedovic and Sivaganesan, 2002; Teh, 2010).

2.1.3 Implementation of Dirichlet Processes

The important characteristic of the Dirichlet process is that samples from the DP are discrete and exhibit clustering effects, which is useful for grouping objects together. The realizations of samples from the Dirichlet process often can be described by the Polya urn process (Blackwell and MacQueen, 1973), the Stick-Breaking Process (Sethuraman, 1991), and the Chinese Restaurant Process (Aldous, 1985). In this section, we review the Chinese restaurant process and the stick-breaking process.

Chinese Restaurant Process

A Chinese Restaurant Process (CRP), which is introduced by Aldous (1985), is often used for construing the clustering effects of the samples drawn from the Dirichlet process. Considering a Chinese restaurant with an infinite number of tables, n customers, labeled with $\{1, 2, \ldots, n\}$, come and sit at tables in the restaurant. Starting from the first customer, the first customer occupies the first table, labeled with 1, and the next customer decides whether to sit at the table occupied by the first customer with the probability $\frac{1}{1+\alpha}$ or at a new table with the probability $\frac{\alpha}{1+\alpha}$. Given that $n-1$ customers have already occupied tables, marked as $k = \{1, 2, \ldots, K\}$, then n^{th} customer comes to the restaurant and chooses one of the tables occupied by $n-1$ customers with the probability $\frac{n_k}{n-1+\alpha}$, where n_k is the number of customers already sitting at the table k or an unoccupied table with the probability $\frac{\alpha}{n-1+\alpha}$. Now, a table where n^{th} customer sits is drawn from the following distribution:

$$\begin{aligned} p(n^{th} \text{ customer sits at the table } k \mid 1, 2, \ldots, n-1) &= \frac{n_k}{\alpha + n - 1}, \\ p(n^{th} \text{ customer sits at the new table} \mid 1, 2, \ldots, n-1) &= \frac{\alpha}{\alpha + n - 1}. \end{aligned} \tag{2.3}$$

After we draw the table for n^{th} customer, let K be the total number of tables occupied by n customers. Also, n customers will have a latent variable, which is a table assignment, an integer from 1 to K. We can view the CRP as the random process which partitions n customers into K clusters. Similar to Blackwell-MacQueen urn scheme (Blackwell and Mac-

Queen, 1973), the CRP specifies a distribution over partitions, which are table assignments of n customers. The graphical representation of the Chinese Restaurant Process is displayed in Figure 2.2.

The fact that customers share the same table or sit alone exhibits the clustering effects in CRP. In CRP, each table represents a cluster. Thus, each customer and each table stand for a data point and a cluster, respectively. The total number of tables occupied by customers represents the number of clusters. In addition, Equation 2.3 addresses the important feature of the Dirichlet process, which is that the probability of opening a new table is proportional to a positive real number α.

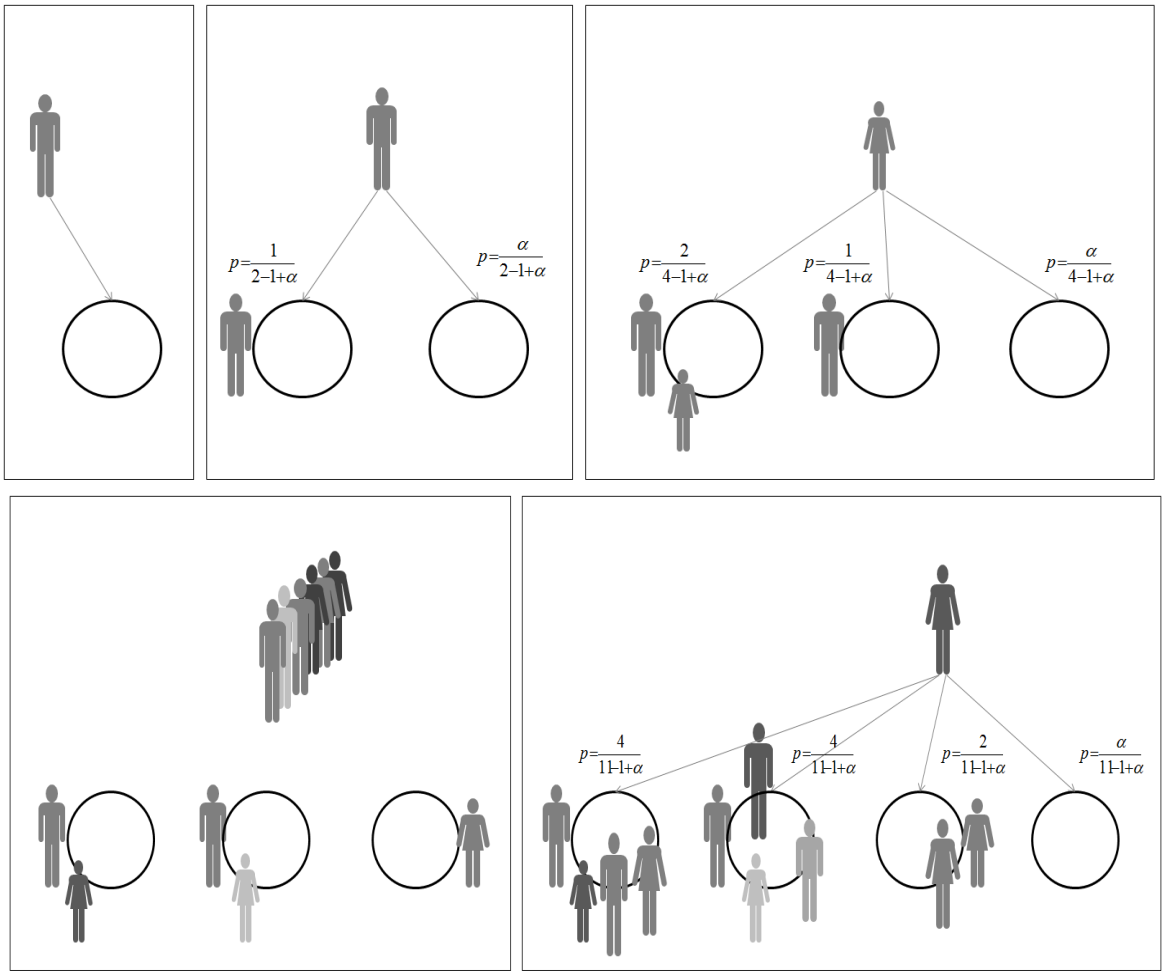


Figure 2.2: A graphical representation of the Chinese Restaurant Process is depicted for describing the clustering effects of the Chinese Restaurant Process.

The Stick-Breaking Process

Previously, we have shown that samples drawn from a Dirichlet process are consisted of the weighted sum of point masses. In stead of a Chinese restaurant process, Sethuraman (1991) has alternatively explained the characteristic of DP samples by using a stick-breaking representation. To construct the Stick-Breaking process, consider $G = \sum_{k=1}^{\infty} \pi_k \delta(\theta = \theta_k)$ where θ is the sample drawn from G_0, and $\sum_{k=1}^{\infty} \pi_k = 1$. Then, assume that we have a stick with length 1 and we draw a random sample β_1 from a Beta distribution with parameters 1 and α. Let $\pi_1 = \beta_1$, which corresponding to the length of the part of the stick we just break off. To obtain π_2, we draw $\beta_2 \sim Beta(a, b)$ and then again let π_2 be $(1 - \beta_1)\beta_2$ which is also the length of the part of stick. Then, we repeat the procedure until $K \to \infty$. For $k = 1, 2, ..., K$, the overall procedure can be written down as:

$$
\begin{aligned}
\beta_k &\sim Beta(1, \alpha), \\
\pi_1 &= \beta_1, \\
\pi_2 &= (1 - \beta_1)\beta_2, \\
&\vdots \\
\pi_K &= \prod_{k=1}^{K-1} (1 - \beta_k)\beta_K.
\end{aligned}
$$

This mechanism can be understood as stick-breaking because a stick with length 1 breaks off with the fraction $\{\pi_1, \pi_2, \ldots, \pi_\infty\}$ and it shows that the concentration parameter α influences the distribution of π. The stick-breaking construction is graphically described in Figure 2.3.

The stick-breaking process is ideal for understanding the realization of Dirichlet process. Also, the stick-breaking process is useful in the Markov chain Monte Carlo algorithms when drawing random samples for the Dirichlet process mixture models because sampling random variables from a Beta distribution is relatively easy to do (Ishwaran and James, 2001).

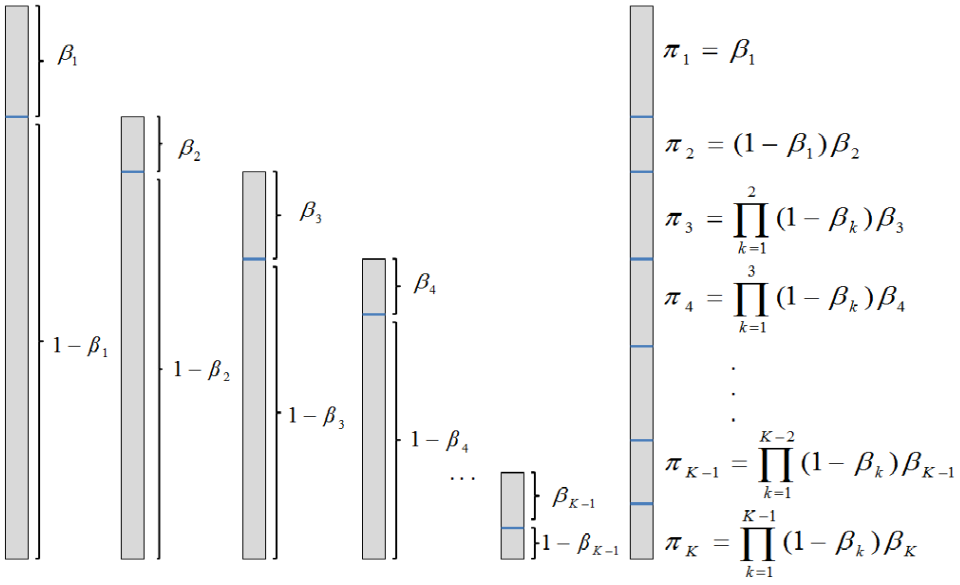


Figure 2.3: A visual depiction of a stick-breaking process. When $K \to \infty$, the stick-breaking process becomes the Dirichlet process.

2.2 Dirichlet Process Mixture Models

As an application of a Dirichlet process prior, a Dirichlet Process Mixture Model (DPMM) has been popularly used as a model-based clustering approach in Bayesian frameworks (Neal, 2000; Figueiredo and Jain, 2002; Medvedovic and Sivaganesan, 2002; Dahl, 2006). Because the Dirichlet process prior in the model enables us to have a various and infinite number of clusters, the DPMM is capable of accommodating an infinite number of clusters and it provides multiple clustering solutions (Medvedovic and Sivaganesan, 2002; Figueiredo and Jain, 2002). In other words, unlike other clustering methods, the DPMM does not require the number of clusters to be fixed in advance when we do not have prior information about it. However, as shown in the previous section, the number of clusters specified by the models turns on a reasonable value of α in the DP, which is application specific.

We formulate the Dirichlet process mixture model by setting up a parametric mixture model. Consider the finite mixture model with K components for a set of n observations $\{y_1, y_2, \ldots, y_n\}$:

$$\begin{aligned} \theta_1, \theta_2, \ldots, \theta_K &\sim G, \\ \pi_1, \pi_2, \ldots, \pi_K &\sim Dir(\frac{\alpha}{K} J), \\ c_i \mid \boldsymbol{\pi} &\sim Mult(1, \boldsymbol{\pi}), \\ y_i \mid c_i, \theta_1, \theta_2, \ldots, \theta_K &\sim F(\theta_{c_i}), \end{aligned} \tag{2.4}$$

where $i = 1, 2, \ldots, n$, $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_K\}$ is a parameter vector, G is a probability distribution, $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots, \pi_K\}$ is the mixing proportion for K components, $J = \mathbf{1}_{1\times K}$, and c_i for y_i is the component label ranged from 1 to K, which indicates the cluster assignment for the observation y_i. Probability distribution F with parameter $\theta = \{\theta_1, \ldots, \theta_K\}$ is underlying distribution for observations y. In semi-parametric Bayesian approach, G in Equation 2.4 is unknown, and we assign the Dirichlet process prior for G. Then, the DPMM is formally

represented as:

$$\begin{aligned} G \mid \alpha, G_0 &\sim DP(\alpha, G_0), \\ \theta_1, \theta_2, \ldots, \theta_n &\sim G, \\ y_i | c_i, \theta_1, \theta_2, \ldots, \theta_n &\sim F(\theta_{c_i}), \end{aligned} \tag{2.5}$$

where $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_n\}$ is the collection of the latent parameters which demonstrates clustering effects of the Dirichlet process prior. The number of clusters is automatically estimated by the number of unique $\boldsymbol{\theta}$, and the estimation of $\boldsymbol{\theta}$ tells us the cluster assignments which distribution the data points are come from. For example, data points labeled with c_3 belong to cluster c_3 and these data points share the same parameter θ_{c_3}. Also, if a data point is newly observed, the DP in the DPMM can allow the new data point either to belong to the existing clusters or to declare its own cluster. The DPMM in Equation 2.5 is visualized in Figure 2.4.

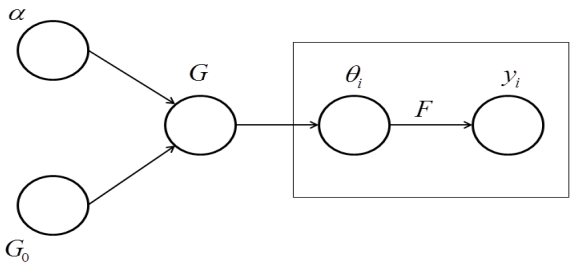


Figure 2.4: The graphical representation of the Dirichlet process mixture model.

The DPMM not only considers the data likelihood as a mixture model, but it also combines the clustering effect by assigning the DP prior over the latent parameters. Thus, the actual number of clusters in DPMM is not fixed and automatically estimated by modeling data according to the nature of DP. The DPMM is a compelling application for clustering analysis in order to alternate finite mixture models with model selection procedures and ad-hoc clus-

tering approaches such as the K-means clustering approach and agglomerative hierarchical clustering algorithms.

2.3 Markov Chain Monte Carlo Algorithms for Dirichlet Process Mixture Models

Due to the advances in Markov Chain Monte Carlo (MCMC) algorithms, sampling from the posterior distribution of parameters, which follow the Dirichlet process, has become computationally feasible (Neal, 2000). In this section, we summarize useful MCMC algorithms to draw random samples from the posterior distribution of Dirichlet process mixture models, including two algorithms that are reviewed and summarized in Neal (2000). The notations for these MCMC algorithms are provided in Table 2.1.

Collapsed Gibbs sampling

The collapsed Gibbs sampling algorithm for the DPMM requires us to choose the base measure G_0 to be conjugate to the distribution f (Escobar, 1994; MacEachern, 1994; Neal, 2000). Let the state of the Markov chain consist of $\boldsymbol{c} = (c_1, c_2, \ldots, c_n)$ and $\boldsymbol{\theta} = (\theta_c : c \in$

Table 2.1: Notations for MCMC algorithms summarized by Neal (2000) are provided.

Notation	
$\boldsymbol{y} = \{y_1, \ldots, y_n\}$	data points with the size of n
y_i	i^{th} data point
c_i	a class indicator associated with y_i
c_{-i}	c_j for $j \neq i$
$\boldsymbol{\theta} = \{\theta_{c_1}, \ldots, \theta_{c_n}\}$	a latent parameter corresponding to class indicator c for y
$n_{-i,c}$	the number of c_j for $j \neq i$ that are equal to c
f	a probability distribution of y
G_0	a base measure in Dirichlet process
α	a concentration parameter in Dirichlet process

$\{c_1, c_2, \ldots, c_n\}$) For $i = \{1, 2, \ldots, n\}$, if no observations are labeled with c_i ($n_{-i,c_i} = 0$), remove θ_{c_i} from the present state of MCMC and sample the new value for c_i with the following probabilities (Neal, 2000):

$$\begin{aligned} P(c_i = c \,|c_{-i},\, y_i,\, \boldsymbol{\theta}) &\propto \frac{n_{-i,c}}{n-1+\alpha} f(y_i|\theta_c), \\ P(c_i \neq c \,|c_{-i},\, y_i,\, \boldsymbol{\theta}) &\propto \frac{n_{-i,c}}{n-1+\alpha} \int f(y_i|\theta) dG_0(\theta), \end{aligned} \tag{2.6}$$

where $n_{-i,c}$ is the number of c_i for $i \neq j$. Then, $\boldsymbol{\theta}$ is sampled from the posterior distribution of $\boldsymbol{\theta}$. This sampling algorithm is feasible when we are able to compute $\int f(y_i|\boldsymbol{\theta}) dG_0(\boldsymbol{\theta})$ and to draw samples from the posterior distribution of $\boldsymbol{\theta}$. This implies that the base measure G_0 is required to be conjugate to a distribution f.

Gibbs sampling with an auxiliary variable

We briefly describe a Gibbs sampling with an auxiliary variable in Neal (2000), which addresses the case that G_0 is not conjugate to f and uses auxiliary variables corresponding to class indicators. Let c_i and k^- denote the latent class variable and the number of distinct c_i for $i \neq j$, respectively. Suppose that $h = k^- + m$, where m is an arbitrary natural number. Before sampling the parameter θ_c, we label c_i with the values in $\{1, 2, 3, \ldots, k^-\}$. For the case that $c_i \neq c_j$, we label c_i with $k^- + 1$ and draw θ_c for $k_- + 1 < c \leq h$. Then, we sample a new value for c_i from $\{1, 2, ..., h\}$ with the following probabilities (Neal, 2000):

$$P(c_i = c \,|c_{-i},\, y_i,\, \theta_1,\, \ldots,\, \theta_h) \propto \begin{cases} \frac{n_{-i,c}}{n-1+\alpha} f(y_i|\theta_c) & \text{for } i \leq c \leq k^- \\ \frac{\alpha/m}{n-1+\alpha} f(y_i|\theta_c) & \text{for } k^- < c \leq h \end{cases}, \tag{2.7}$$

where $n_{-i,c}$ is the number of c_i for $i \neq j$. This algorithm does not require the base measure G_0 to be conjugate to f. More MCMC algorithms for sampling in the DPMM are well introduced in Neal (2000).

Chapter 3

Linkage Based Dirichlet Processes: Methodology

The fact that the number of clusters relies on the concentration parameter in the Dirichlet process prior emphasizes the importance of estimation of the concentration parameter (α) for gaining the adequate number of clusters given observed data. This chapter is divided into the following sections. Section 3.1 reviews related work for estimating the concentration parameter. In Section 3.2, we propose the new method, which we will call linkage based Dirichlet process.

3.1 Related Work

In clustering analyses, the most important issue is related to defining an appropriate number of clusters given observed data (Fraley and Raftery, 1998). The Dirichlet process prior demonstrates a "*rich-gets-richer*" property that the probability of choosing the table in the CRP is proportional to the number of customers who already sat at the table (Pitman et al., 2002). Given the fixed number of customers in the CRP, the probability of obtaining a new

cluster is proportional to the size of α. We have discussed these properties in Chapter 2. Antoniak (1974) proves that the distribution of the number of clusters is dependent on both the number of observations in the data and the concentration parameter in that

$$p(K \mid \alpha, n) = c_n(K) n! \alpha^K \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)},$$

where K is the number of clusters and $c_n(K)$ is Stirling number (Antoniak, 1974). When the number of observations $n \to \infty$, the expected number of clusters K for observations is:

$$E(K|\alpha, n) = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1} \approx \alpha log(n), \tag{3.1}$$

which was established by Liu (1996). Equation 3.1 certainly demonstrates that the expected number of clusters relies on the size of α and n. With the fixed n observations, the expected number of clusters increases as the size of α increases. Thus, depending on the size of the concentration parameter, the Dirichlet process mixture model may either overestimate the number of clusters or underestimate it. This emphasizes the importance of estimating the concentration parameter to obtain the appropriate number of clusters.

Due to the pioneering of work by West (1992), a Gamma prior with a shape parameter a and a scale parameter b is assigned over the concentration parameter. Then, West (1992) induces the resulting posterior distribution of the concentration parameter, which is also a Gamma distribution, so that a Gibbs sampling can draw samples for estimating the concentration parameter. However, this approach gives rise to estimating the hyperparameters a and b in the Gamma prior. Related to the work of West (1992), Dorazio (2009) presents the means of estimating the hyperparameters, a and b, in the Gamma prior in the Dirichlet process through *KL-divergence*. Accordingly, West (1992), Escobar and West (1995), Liu (1996), Dorazio (2009), McAuliffe et al. (2006), and Rabaoui et al. (2011) discuss how to estimate the concentration parameter α in the DP. Liu (1996) demonstrates the maximum likelihood estimate of the concentration parameter α, which satisfies Equation 3.1. Liu

(1996) encourages the use of an empirical Bayes approach for making inferences about the concentration parameter.

McAuliffe et al. (2006) introduces an approach for estimating both the base distribution G_0 and the concentration parameter α. McAuliffe et al. (2006) posits that an estimation of the base measure G_0 is important when the true density of the $\theta_i's$ is skewed and does not follow the parametric form of the probability densities in the exponential family. They suggest employing both a kernel density estimation and an empirical Bayes method to estimate G_0. The constructed kernel density estimate ($\hat{G_0}$) for G_0 in McAuliffe et al. (2006) is:

$$\begin{aligned} E(\hat{G_0^*}|y_1, ..., y_n) &= E\bigg(\frac{1}{K(\theta_{1:n})}\sum_{i=1}^{K(\theta_{1:n})}\kappa_h(\theta_i^*|y_{1:n})\bigg) \\ &\approx \frac{1}{B}\sum_{b=1}^{B}\bigg(\frac{1}{K_b}\sum_{i=1}^{K_b}\kappa_{h_b}(\theta_i^{b*})\bigg) = \hat{G_0}, \end{aligned} \tag{3.2}$$

where B is a specified number of previous MCMC samples. $K(\theta_{1:n})$ is the number of unique θ_i^*'s observed, κ_h is the kernel with width (h) for density estimation, and K_b is the number of unique $\theta_{1:K_b}^{b*}$. After estimating G_0 in Equation 3.2, they draw the θ_i^*'s from the point estimate $\hat{G_0}$. Then, the concentration parameter is estimated by solving the following equation:

$$\sum_{i=1}^{n}\frac{\alpha}{\alpha+i-1} \approx \frac{1}{B}\sum_{b=1}^{B}K_b. \tag{3.3}$$

Rugging in these estimates, at each iteration of a Gibbs sampler, constitutes an empirical Bayesian method for specifying $\hat{G_0}$ and $\hat{\alpha}$ (McAuliffe et al., 2006).

However, with respect to the estimation of the concentration parameter α, the method proposed by McAuliffe et al. (2006) presents several possible limitations of their method. First of all, the estimated base measure is a weighted average of the samples from an MCMC

procedure via a Kernel Density Estimation and an empirical Bayes method, rather than a sample itself from the MCMC procedure. This implies that the estimated concentration parameter is no longer a sample from the MCMC and does not hold the Markov chain property. Secondly, their method has the potential to get trapped in inappropriate clusters. Once they find the membership of observations, they update the concentration parameter using only the number of observations and the number of defined clusters in the MCMC algorithm.

In summary, the methods mentioned in Section 3.1 for estimating the concentration parameter have the common limitation of relying on only the number of observations which are assigned to the clusters. In addition, these methods do not account for the shape of clusters and the distances between clusters. We thus emphasize the necessity for the new method, which incorporates not only the size of n and α, but also the linkages between defined clusters.

3.2 Linkage Based Dirichlet Processes

In this section, we propose a new method for estimating the concentration parameter in Dirichlet processes. The estimation of the concentration parameter is based on the linkages between clusters. Before we propose our method in detail, we review several meaningful metrics that measure distances between clusters in probability scales, and an empirical Bayes approach.

3.2.1 Meaningful Metrics for Measuring Distances between Clusters

In statistics, measuring distances between clusters is indispensable for clustering analyses. For instance, agglomerative clustering analyses use several linkage functions, such as complete linkage, average linkage, and single linkage, for measuring distances between two clusters. The range of distances between a pair of clusters is between 0 and ∞. Thus, it is challenging to determine whether two clusters are close enough to be merged or separated. In addition, given observed data points $(x_1, x_2, \ldots, x_n)$, the K-means clustering method employs the metric which is called *Within-cluster Sum of Squares* (WSS) and determines clustering assignments by minimizing it:

$$WSS = \arg\min_{\mathbf{S}} \sum_{k=1}^{K} \sum_{\mathbf{x} \in S_k} \|\mathbf{x} - \boldsymbol{\mu}_k\|^2 ,$$

where $\mathbf{S} = \{S_1, S_2, \cdots, S_K\}$ represents clusters, K is the number of clusters, and μ_k is the mean of observations assigned to S_k. However, by WSS, it is somewhat problematic to determine how close the clusters are. We expect proximity between clusters to be stretched, since our goal is to define well-separated clusters via any clustering algorithms.

In model-based clustering algorithms, formal probability mixture models assume that observations within a cluster follow a distribution f with parameter θ. This means that each cluster has its own distribution. With respect to explaining proximity between clusters, some metrics which quantify distances between probability distributions are appropriate. Commonly used probability measures include the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951; Kullback, 1987), the earth mover's distance (Rubner et al., 2000; Levina and Bickel, 2001), the Hellinger distance (Hellinger, 1909; Bhattacharyya, 1946; Nikulin, 2001), the total variation distance (Dunford and Schwartz, 1958), and so on.

In order to account for proximity between K clusters, we measure distances by a metric, which range is between 0 and 1, for all $\binom{K}{2}$ pairs. If K clusters are well separated, each cluster has a distinct probability density. Then, distances between any pair of clusters are approximately 1. On the other hand, the distance between two clusters, which are almost completely overlapping, is close to 0. Any metric, which gives results between 0 and 1, is suitable for our proposed method. Also, metrics should satisfy the properties of a distance measure as follows:

The properties of a distance measure

A metric d on a set X such that $d : X \times X \to [0, \infty)$ is called a distance function if the following conditions are satisfied for all x, y, z in X:

1. Non-negativity: $d(x, y) \geq 0$,
2. Identity of indiscernibles: $d(x, y) = 0$ if and only if $x = y$,
3. Symmetry: $d(x, y) = d(y, x)$,
4. Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$.

Now we introduce possible distance metrics, which give us an insight into the proximity between clusters with their range. For notational convenience, we assume that f and g are probability measures on a σ-algebra $\mathcal{F}$ of subsets of the sample space Ω and $x \in \Omega$.

Total variation distance

Total variation distance is a distance measure that calculates the difference between two probability distributions. In Bayesian statistics, this measure is often used for assessing the convergence of the MCMC chain to the stationary distribution from the current distribution (Reutter and Johnson, 1995). The mathematical expression of total variation distance

between two probability measures is as follows:

$$\begin{aligned} TVD(f,g) &= \sup_{x\in\mathcal{F}} |f(x) - g(x)| \\ &= \int_{f>g} \Big(f(x) - g(x) \Big) \mathrm{d}x \\ &\approx \frac{1}{2} \sum_{i=1}^{n} |f(x_i) - g(x_i)|. \end{aligned}$$

Operating under the assumption that clusters have their own distributions, total variation distance measures the distance between two probability measures. A total variation distance of 0 suggests that two probability distributions or two clusters completely overlap while a total variation distance of 1 means two are very distinct. We employ total variation distance in order to measure the distance between clusters.

Hellinger distance

Hellinger distance introduced in Bhattacharyya (1946) is a metric for measuring the distance between two probability density functions f and g. The squared Hellinger distance with an integral is expressed as

$$\begin{aligned} HD^2(f,g) &= \frac{1}{2} \int \Big(\sqrt{f(x)} - \sqrt{g(x)} \Big)^2 dx \\ &= 1 - \int \sqrt{f(x)g(x)}\, dx. \end{aligned}$$

When estimating a density or clustering observations, we use observations which are discrete and are assumed to be from the probability distribution. Thus, Hellinger distance between two continuous probability density functions can be approximated by the following equation:

$$\begin{aligned} HD(f,g) &= \frac{1}{\sqrt{2}} \left(\int \left(\sqrt{f(x)} - \sqrt{g(x)} \right)^2 \mathrm{d}x \right)^{1/2} \\ &\approx \frac{1}{\sqrt{2}} \left(\sum_{i=1}^{n} \left(\sqrt{f(x_i)} - \sqrt{g(x_i)} \right)^2 \right)^{1/2}. \end{aligned}$$

Hellinger distance and total variation distance for two probability distributions have the following relationship: $H^2(f,g) \leq TVD(f,g) \leq \sqrt{2}H(f,g)$ (Gibbs and Su (2002)). For two multivariate normal distributions $f \equiv \mathcal{N}(\mu_1, \Sigma_1)$ and $g \equiv \mathcal{N}(\mu_2, \Sigma_2)$, squared Hellinger distance is obtained as:

$$HD^2(f,g) = 1 - \frac{\det(\Sigma_1)^{1/4}\det(\Sigma_2)^{1/4}}{\det\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{1/2}} \exp\left\{-\frac{1}{8}(\mu_1-\mu_2)^T\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{-1}(\mu_1-\mu_2)\right\}. \tag{3.4}$$

The Hellinger distance in Equation 3.4 is easier to compute than total variation distance when clustering multidimensional data by a linkage based Dirichlet process mixture model.

Cosine divergence

Cosine divergence is one of metric among J-divergence (Chung et al., 1989). The cosine divergence can be written as follows:

$$\begin{aligned} CD_{n,\alpha}(f,g) &= \frac{1}{2}\left[1 - \int (f(x)g(x))^{1/2} cos\left(s \cdot log_2 \frac{f(x)}{g(x)}\right)\right] \\ &\approx \frac{1}{2}\left[1 - \sum_{i=1}^{n} (f(x_i)g(x_i))^{1/2} cos\left(s \cdot log_2 \frac{f(x_i)}{g(x_i)}\right)\right], \end{aligned} \tag{3.5}$$

where s is a degree in J-divergence, which effciently adjusts the logarithm base measuring the divergence between f and g. Cosine divergence has useful properties in that this metric is symmetric and bounded from 0 to 1. If and only if two distributions are same, $CD_{n,s}$ becomes 0.

Jensen-Shannon divergence

In statistics, Jensen-Shannon divergence also measures the proximity between two probability distributions. This metric is also known as total divergence to the average (Dagan et al., 1997). This metric is bounded between 0 and 1. Thus, it will provide the clear criterion to

decide the similarity between clusters. It is defined by

$$\begin{aligned} \mathrm{JSD}(f,g) &= \frac{1}{2}\Bigg[\int f(x) log \frac{f(x)}{\frac{1}{2}(f(x)+g(x))}\mathrm{d}x + \int g(x) log \frac{g(x)}{\frac{1}{2}(f(x)+g(x))}\mathrm{d}x\Bigg] \\ &\approx \frac{1}{2}\Bigg[\sum_{i=1}^{n} f(x_i) log \frac{f(x_i)}{\frac{1}{2}(f(x_i)+g(x_i))} + \sum_{i=1}^{n} g(x_i) log \frac{g(x_i)}{\frac{1}{2}(f(x_i)+g(x_i))}\Bigg]. \end{aligned} \tag{3.6}$$

Jensen-Shannon divergence is suggested in order to overcome the drawback of Kullback-Leibler divergence, which is not symmetric. We can view Jensen-Shannon divergence as a symmetrized and smoothed version of KullbackLeibler divergence.

Linkage functions

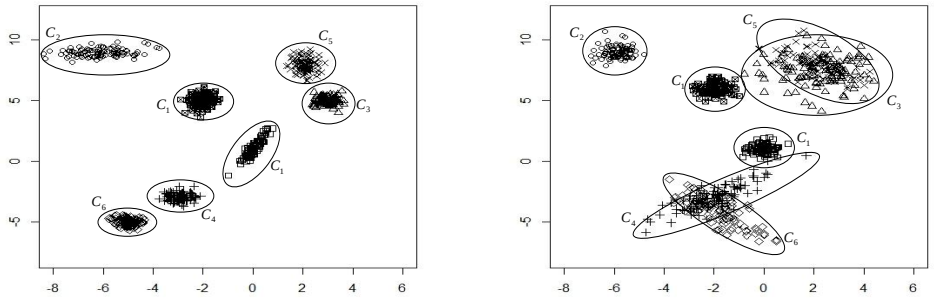
Alternatively, when computing exact probability distance between clusters is difficult, we can use linkage functions in agglomerative hierarchical clustering methods. The linkage functions can be scaled to the value between 0 and 1 by employing the following equation:

$$dist(i,j) = \frac{d(i,j) - d_{min}}{d_{max} - d_{min}}, \tag{3.7}$$

where $d(i,j)$ is the distance between cluster i and cluster j calculated by the linkage functions such as the complete linkage, the average linkage, and average linkage function. d_{min} and d_{max} are the minimum and maximum among distances between pair-wised clusters, respectively. $dist(i,j)$ ranges between 0 and 1, and 1 of $dist(i,j)$ indicates two clusters are distant. We can expect that using linkage functions and scaling the distances are useful when the data is on the high dimensions.

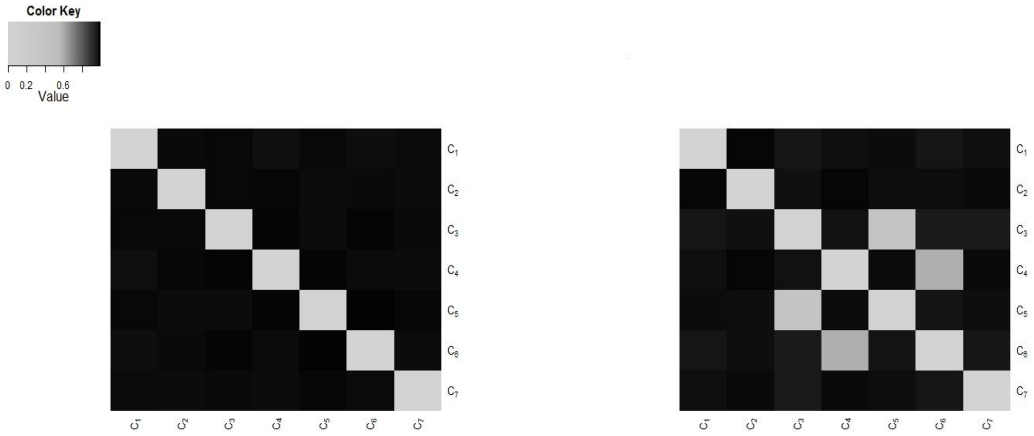
We have listed distance metrics which can be used in our proposed method. Since clustering analyses deal with high dimensional data in general, we strongly suggest utilizing Hellinger distance or linkage functions. We employ the concept of distance measures in order to measure the distance between clusters. Then, distance measures in this section allow for us to not only make inferences about the concentration parameter, but also accurately

identify the number of clusters in the DPMM. Figure 3.1 illustrates two simulated data from a Gaussian mixture distribution with 7 components. The left plot in Figure 3.1 depicts 7 non-overlapping clusters while the right one in Figure 3.1 plots 7 clusters, some of which are overlapping. Intuitively, the TVD between pairwise clusters on the left plot in Figure 3.1 would be approximately 1. On the other hand, the TVD between overlapping clusters on the right one in Figure 3.1 would be far less than 1. Figure 3.2 shows the proximity between pairwise clusters in Figure 3.1 for two different scenarios by calculating the total variation distance.



(a) Non-overlapping clusters (b) Overlapping clusters

Figure 3.1: Panel (a) shows the 7 non-overlapping clusters and Panel (b) shows the overlapping clusters.



(a) TVD for non-overlapping clusters (b) TVD for overlapping clusters

Figure 3.2: Panel (a) and (b) describe the total variation distances for all possible pairwise clusters in Figure 3.1 (a) and (b).

3.2.2 Empirical Bayes Methods

We briefly review the class of empirical Bayes methods, because part of our proposed method utilizes the empirical Bayes method when estimating a concentration parameter. The empirical Bayes approach is an application for parameter estimation and inference. The empirical Bayes approach is placed between classical and Bayesian methods because this approach borrows pieces from each. In general, an estimation step is done by classical techniques and an inference step is completed by Bayesian techniques in general (Casella, 1992). We provide an example that describes how the empirical Bayes approach mixes classical and Bayesian methods. Consider the hierarchical model, i.e., data likelihood and prior:

$$x \sim f(x|\theta)$$
$$\theta \sim g(\theta|\lambda),$$

where λ is a hyperparameter in a prior distribution of θ and then $p(x|\lambda) = \int f(x|\theta)g(\theta|\lambda)d\theta$. In the empirical Bayes approach, instead of specifying λ, we obtain $\hat{\lambda}$ based on $p(x|\lambda)$ by using classical methods such as method of moments or maximum likelihood estimation in order to estimate λ. Then, we substitute λ with $\hat{\lambda}$ so that we have the posterior distribution $p(\theta|x, \hat{\lambda})$. Because the empirical Bayes method uses observed data to specify the prior specification for λ and utilizes classical methods for estimation, the empirical Bayes approach is not the fully Bayesian technique. Also, since we substitute λ with $\hat{\lambda}$ for the posterior distribution $p(\theta|x, \hat{\lambda})$, the empirical Bayes approach is not the fully classical technique. In other words, instead of specifying unknown hyperparameter in fully Bayesian techniques, the empirical Bayes approach attempts to estimate unknown hyperparameter and substitutes estimated hyperparameter into Bayesian quantity. Thus, empirical Bayes approach is a hybrid application of both classical and Bayesian techniques (Casella, 1992).

For easier computation in the empirical Bayes approach, in order to estimate the hyperparameter in the prior distribution, Casella (2001) suggests the empirical Bayes Gibbs sam-

pling, which uses samples from Gibbs sampler for estimation. Both McAuliffe et al. (2006) and our proposed method use MCMC samples for estimating the concentration parameter α instead of assigning a prior distribution for this parameter.

3.2.3 Linkage Based Dirichlet Processes

Let D denote the distance matrix containing the distances between $\binom{K}{2}$ pairs of clusters. Then, $D(C_i, C_j)$ denotes the distance between clusters C_i and C_j for i and $j \in \{1, 2, \ldots, K\}$. The sum of the upper triangular elements of D for K non-overlapping clusters, $\sum_{i<j} D(C_i, C_j)$, is approximately $\binom{K}{2}$, where the range of $D(C_i, C_j)$ is between 0 and 1. Consider the following function for K distinct clusters:

$$S(C, K) = \binom{K}{2} - \sum_{i<j}^{K} D(C_i, C_j), \tag{3.8}$$

where C represents clusters $(C_1, C_2, ..., C_K)$. The maximum value which $\sum_{i<j}^{K} D(C_i, C_j)$ can have is $\binom{K}{2}$. Thus, it is apparently clear that the function $S(C, K)$ is greater than or equal to 0. $S(C, K)$ is approximately 0 when all clusters are accurately defined and clearly distinguishable. Thus, the general idea behind the use of the function $S(C, K)$ is to validate that K specified clusters is an ideal number of clusters for observations in terms of probability distances between clusters. We use this idea to calibrate the concentration parameter α in the DP prior.

Given data, let K' be an ideal number of clusters, such that $\binom{K'}{2} = \sum_{i<j}^{K} D(C_i, C_j)$. K' represents the appropriate number of clusters, at least with respect to the distance measure, which is a useful metric for suggesting whether two structures should be combined or not. Our goal is to search for K' when the DPMM defines K clusters and the distances between K clusters are calculated. Then, the positive solution for $\binom{K'}{2} = \sum_{i<j}^{K} D(C_i, C_j)$ is derived

in Equation 3.10,

$$\frac{K'^2 - K'}{2} = \sum_{i<j}^{K} D(C_i, C_j), \tag{3.9}$$

which yields:

$$K' = \frac{1 + (1 + 8\sum_{i<j}^{K} D(C_i, C_j))^{1/2}}{2}. \tag{3.10}$$

Note that K in the distance matrix D is the number of clusters defined via the DPMM. K' is an appropriate number of clusters which reflects the probability distances between K clusters. For our proposed method, we find K' in Equation 3.10 and use it for estimating the concentration parameter.

Liu (1996) proves that the expected number of clusters is conditional on both the number of observations and the concentration parameter α as shown in Equation 3.1. For estimating the concentration parameter, we replace $E(K \mid \alpha, n)$ with K' in Equation 3.10. Then, we rewrite Equation 3.1 as follows:

$$K' = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1}. \tag{3.11}$$

For updating the concentration parameter in MCMC chain, we solve Equation 3.11 with respect to α. We use Brent's method (Brent, 1973), which borrows benefits of both bisection method and Newton's method as a root-finding algorithm, since Equation 3.11 is not a closed linear form of α. Then, we estimate the concentration parameter α which satisfies Equation 3.11. Equation 3.11 is the same equation in the empirical Bayes approach introduced by McAuliffe et al. (2006) and Liu (1996) except K'. Recall that McAuliffe et al. (2006) uses B samples from previous draws in order to estimate the concentration parameter, and their estimators for a concentration parameter are no longer MCMC samples but are weighted averages of MCMC samples at each iteration. However, our method uses random sample from the present state of MCMC chain in order to hold Markov chain property for estimating the concentration parameter.

3.2.4 Gibbs Sampling Implementation

Our proposed method can be applied into any MCMC techniques, which draw samples from the posterior distribution of θ in Figure 2.4. We introduce how to incorporate our proposed method in the Gibbs sampler.

Suppose we have n observations and fit the DPMM for clustering. First, we sample the latent class variable $\boldsymbol{c} = \{c_1, c_2, \ldots, c_{n-1}, c_n\}$ corresponding to clustering memberships. Let $\boldsymbol{c}^* = \{c_1, c_2, \ldots, c_K\}$ be the unique value of sampled $\boldsymbol{c}$. Then, with partitioned n observations into K defined clusters based on their sampled latent class variables, we measure the probability distance between K clusters by using the kernel density estimation and construct the distance matrix ($\boldsymbol{D}$). After finding $\boldsymbol{D}$, we obtain the solution K' in Equation 3.10 and estimate α corresponding to K'. Our proposed method can be applied to any MCMC techniques for drawing samples from the posterior distribution of θ. We summarize the procedure for implementing our approach within MCMC techniques in Algorithm 1. This method yields the estimate for the concentration parameter, which corresponds to the optimal number of clusters for observations.

Algorithm 1 MCMC algorithm for LB-DP mixture models

Initialize all parameters.
for $t = 1$ to T **do**
  for $i = 1$ to n **do**
    Draw cluster indicator $c_i^{(t)}$ for $i = 1, 2, \ldots, n$ with

$$p(c_i = c|.) = \pi_c^{(t-1)} \prod p(y_i \mid \theta_c^{(t-1)}).$$

  end for i
  Let $c^* = 1, 2, \ldots, K^{(t)}$ be the unique value of $c^{(t)}$.
  Sample $u_c \sim beta(1 + n_c, \alpha^{(t-1)} + \sum_{c+1}^{C} n_s)$ and update $\pi_c = u_c \prod_{s=1}^{c-1}(1 - u_s)$.
  Draw a new sample for $\theta_{c^*}^{(t)}$ from $p(\theta_{c^*}^{(t)} \mid y_{c^*})$.
  Construct the distance matrix $D^{(t)}$ for $K^{(t)}$ clusters.
  Update $K'^{(t)}$ by solving

$$K'^2 - K' - 2 \sum_{\substack{k<j, \\ k,j=1,\ldots,K^{(t)}}} D^{(t)} = 0.$$

  Update $\alpha^{(t)}$ by solving

$$\sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1} = K'^{(t)}.$$

end for t

Chapter 4

Linkage Based Dirichlet Processes: Simulation and Comparisons

In this chapter, we present a sequence of simulation studies for comparing the performance of linkage based Dirichlet processes under varying conditions. The rest of the chapter is organized as follows. Section 4.1 presents our simulation designs. Then, Section 4.2 provides simulation results corresponding to our simulation designs and demonstrates the performance of the linkage based Dirichlet processes against the method suggested by McAuliffe et al. (2006), which we refer to as DP-MBJ. For simulation studies in Chapter 4, except the mixture of univariate Gaussian example in section 4.1.1, we have truncated $K = 10$ for efficient computation. Also, in our simulation studies except the simulation design in Section 4.1.1, we also assume that $\sigma_i \sim DP$ or $\Sigma_i \sim DP$ where $i = \{1, 2, \ldots, n\}$ given n observations. Then, we conclude our findings in Section 4.3.

4.1 Simulation Design

In this section, we introduce our simulation designs for generating data points: a mixture of univariate Gaussian distributions, an overlapping/non-overlapping mixture of bivariate Gaussian distributions, and an overlapping/non-overlapping mixture of 5-D multivariate Gaussian distributions.

4.1.1 Mixture of Univariate Gaussians

We generate $n = 1000$ observations from a Gaussian mixture distribution with a mean vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_{1000})$, which are samples drawn from the Dirichlet process, and $\sigma = 0.1$ fixed. The simulated observations y_i are from the following distribution:

$$y_i \sim N(\theta_i, \sigma^2),$$
$$\theta_i \sim G,$$
$$G \sim DP(\alpha, G_0),$$

where $i = \{1, 2, \ldots, 1000\}$, and G_0 is the normal distribution with $\mu = 0$ and $\sigma = 3$. The concentration parameter used for generating $\boldsymbol{\theta}$ from the DP is diversely chosen in order to produce various numbers of clusters. The number of simulated clusters K is from 6 to 20. For each K, we have 50 simulated datasets. To apply DP-MBJ into simulated observations, we use $B = 100$ samples to estimate the base measure G_0 and the concentration parameter α. Given our simulated data, we aim to find the optimal number of clusters, and estimate α.

4.1.2 Mixture of Bivariate Guassians

Our simulation is based on K-component mixture of bivariate Gaussian distributions, where $K = \{3, 4, 5, 6\}$. Each component weight is $1/K$. We have two schemes to generate the mixture of bivariate Guassian distributions: non-overlapping clusters and overlapping clusters.

Non-overlapping clusters

Our simulation design to simulate n data points from K non-overlapping clusters is as follows:

Step 1. Simulate a mean vector $\boldsymbol{\theta_k} = (\theta_1, \theta_2)$ and a covariance matrix, $\boldsymbol{\Sigma_k} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$.

Step 2. Generate a cluster $Y_k \sim MVN(\boldsymbol{\theta_k}, \boldsymbol{\Sigma_k})$.

Step 3. Repeat step $1 \sim 3$ until we have K clusters.

We choose mean vectors to make sure that we have K distinct clusters.

Overlapping clusters

We show how we generate a pair of overlapping clusters as follows:

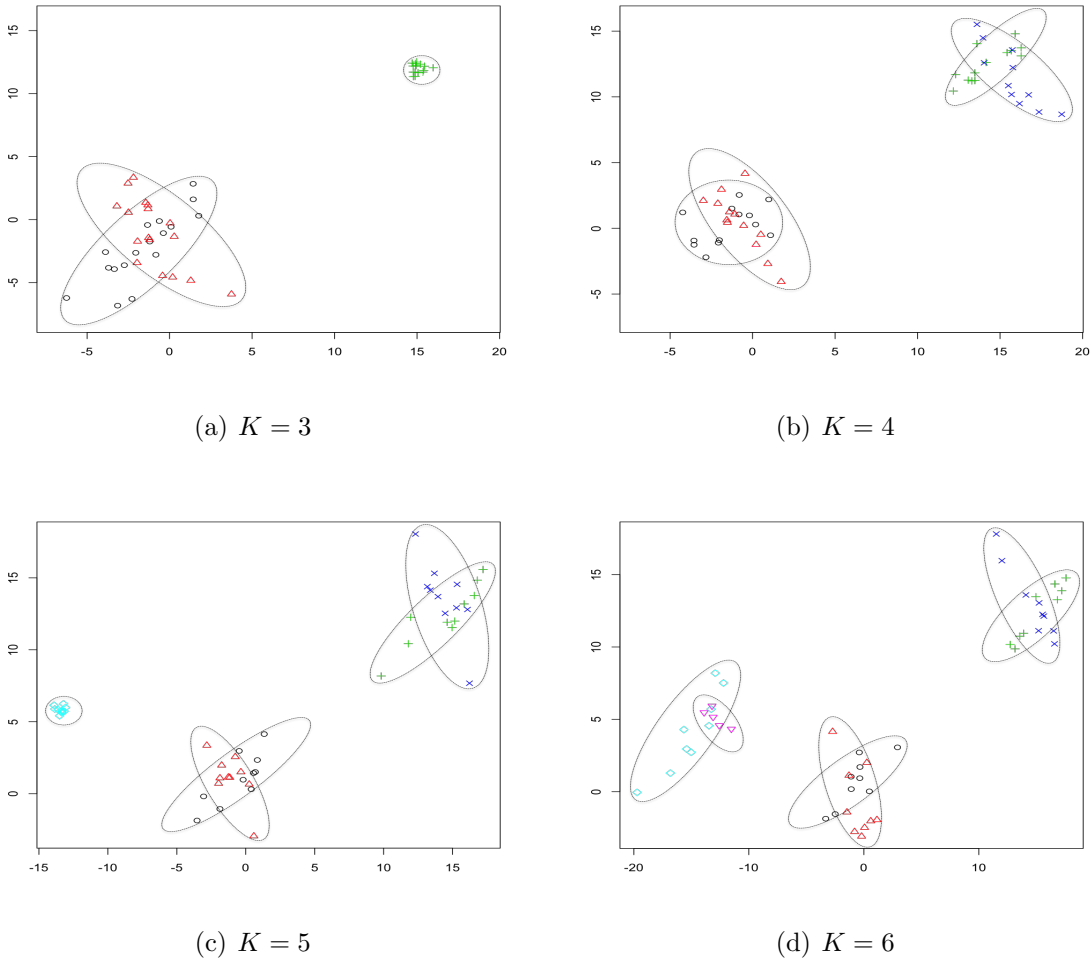
Step 1. Simulate a mean vector $\boldsymbol{\theta} = (\theta_1, \theta_2)$ and a covariance matrix, $\boldsymbol{\Sigma_x} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$.

Step 2. Choose τ and rotate the covariance matrix $\boldsymbol{\Sigma_x}$ by using the rotation matrix $\boldsymbol{R} = \begin{pmatrix} cos(\tau) & -sin(\tau) \\ sin(\tau) & cos(\tau) \end{pmatrix}$. Then, the rotated covariance matrix $\boldsymbol{\Sigma_y} = \boldsymbol{R}\boldsymbol{\Sigma_x}\boldsymbol{R}^T$.

Step 3. Generate two overlapping clusters $X \sim MVN(\boldsymbol{\theta}, \boldsymbol{\Sigma_x})$ and $Y \sim MVN(\boldsymbol{\theta}, \boldsymbol{\Sigma_y})$.

Under this scheme, we generate a pair of overlapping clusters and a distinct cluster for $K = 3$. For $K = 4$, we simulate two pairs of overlapping clusters, and we generate two pairs

of overlapping clusters and a distinct cluster for $K = 5$. For $K = 6$, we have three pairs of simulated overlapping clusters. With $n = 50$, the examples of simulated overlapping clusters for different K are visualized in Figure 4.1. τ is chosen to be in $[\pi/3, 2\pi/3]$ for rotation.



(a) $K = 3$ (b) $K = 4$

(c) $K = 5$ (d) $K = 6$

Figure 4.1: Panels (a), (b), (c), and (d) show $n = 50$ data points from overlapping clusters generated according to the simulation design in Section 4.1.2 for $K = 3, 4, 5$ and 6, respectively.

4.1.3 Mixture of 5-D Multivariate Gaussians

We generate n observations from a mixture of 5-dimensional multivariate Gaussian distributions with a set of mean vectors $\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \ldots, \boldsymbol{\theta_K}$ and a set of covariance matrices $\sum_{\boldsymbol{1}}, \sum_{\boldsymbol{2}}, \ldots, \sum_{\boldsymbol{K}}$ where K is the number of mixture components. The simulation design for the mixture of 5-D multivariate Gaussian distributions is similar to the simulation design in Section 4.1.2.

Non-overlapping clusters

The scheme for generating non-overlapping clusters on 5-D space is similar to the generation of non-overlapping clusters in Section 4.1.2, but the length of the mean vector and the dimension of the covariance matrix are different.

Overlapping clusters

To generate a pair of overlapping clusters on 5-dimensional space, we implement 3D rotation to obtain the rotated covariance matrix. We rotate the first 3 axes, i.e., coordinate rotation of x, y, and z axes are chosen for generating the overlapping clusters (Goldstein, 1965). We present how we generate a pair of overlapping clusters as follows:

Step 1. Simulate a mean vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$.

Step 2. Simulate a covariance matrix, $\boldsymbol{\Sigma_x}$

Step 3. Choose τ and rotate the covariance matrix $\boldsymbol{\Sigma_x}$ by using the rotation matrix

$$\boldsymbol{R} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & cos(\tau) & -sin(\tau) & 0 & 0 \\ 0 & sin(\tau) & cos(\tau) & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} cos(\tau) & 0 & -sin(\tau) & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ sin(\tau) & 0 & cos(\tau) & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} cos(\tau) & -sin(\tau) & 0 & 0 & 0 \\ sin(\tau) & cos(\tau) & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Then, the rotated covariance matrix $\boldsymbol{\Sigma_y} = \boldsymbol{R}\boldsymbol{\Sigma_x}\boldsymbol{R}^T$.

Step 4. Generate two overlapping clusters $X \sim MVN(\boldsymbol{\theta}, \boldsymbol{\Sigma_x})$ and $Y \sim MVN(\boldsymbol{\theta}, \boldsymbol{\Sigma_y})$.

For generating non-overlapping clusters, the mean vectors $\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \ldots, \boldsymbol{\theta_K}$ are chosen to ensure clear separation between the clusters. However, when generating overlapping clusters, two overlapping clusters share the same mean vector. For convenience, let LB-DP:HD and LB-DP:Comp be our proposed method using Hellinger distance and our proposed method using complete linkage function, respectively. Since the construction of $\hat{G}_0$ for high dimensional data in the original DP-MBJ is not easy and causes overestimation of the number of clusters, we leave out the procedure of construction of $\hat{G}_0$ from our simulation studies, except for the mixture of univariate Gaussians.

4.2 Simulation Results

4.2.1 Mixture of Univariate Gaussians

In this section, we aim to find the optimal number of clusters given simulated observations and to estimate α along with these clusters. For convenience, let DP-MBJ and LB-DP be the method proposed by McAuliffe et al. (2006) and our proposed method, respectively. As the range of distances calculated by the complete linkage function is not between 0 and 1, we standardize the distance so that the range of standardized distance is between 0 and 1. The standardized distance $dist(i, j)$ is as follows:

$$dist(i, j) = \frac{d(i, j) - d_{min}}{d_{max} - d_{min}},$$

where $d(i, j)$ is the distance between cluster i and cluster j calculated by complete linkage function. Given simulated observations, let K be the true number of clusters. We also denote K_{opt} the optimal number of clusters, the positive solution in Equation 3.10. Let the ratio

R_{opt} and the ratio R_{true} have the following forms:

$$R_{opt} = \left\| \frac{K_{opt} - \hat{K}}{K_{opt}} \right\|,$$
$$R_{true} = \left\| \frac{K - \hat{K}}{K} \right\|,$$

where $\hat{K}$ is the posterior mode of the number of clusters defined by DP-MBJ or LB-DP. The significance of using the relative ratio R_{opt} and R_{true} is in connection with not only the estimated number of clusters, but also the estimation of the concentration parameter. For example, if R_{opt} or R_{true} is 0, then we think that the estimated number of clusters and $\hat{\alpha}$ are well defined. However, if R_{opt} or R_{true} is substantially large, the estimated number of clusters is far from the optimal number of clusters K_{opt} or the true number of clusters K. Then, we consider the estimated number of clusters and $\hat{\alpha}$ are not appropriate for observed data.

Figure 4.2 and Figure 4.3 illustrate side-by-side boxplots for the distribution of the ratio R_{opt} and R_{true} obtained by LB-DP and DP-MBJ. We refer to LB-DP with total variation distance as LB-DP:TVD, LB-DP with Hellinger distance as LB-DP:HD, and LB-DP with complete linkage function as LB-DP:Comp. If any method works perfectly for the simulated data, then R_{opt} or R_{true} becomes 0. Associated with the estimate of the number of clusters, both Figure 4.2 and Figure 4.3 report that DP-MBJ tends to have higher R_{opt} and R_{true} than our LB-DP:TVD, LB-DP:HD, and LB-DP:Comp as K increases. This implies that our LB-DP is more robust than DP-MBJ when defining the appropriate number of clusters given simulated observations. Also, we can see that LB-DP:TVD, LB-DP:HD, and LB-DP:Comp show similar levels of performance with respect to R_{opt} and R_{true} regardless of the choice of linkage functions.

As we reviewed in Chapter 3, DP-MBJ uses kernel density estimation and the empirical Bayes approach for constructing $\hat{G}_0$, and $\hat{G}_0$ has an effect on defining the proper number of

clusters. Consider the case that we draw samples from $\hat{G}_0$. Because $\hat{G}_0$ is constructed over previously sampled $\theta_{1:n}^{(1)}, \theta_{1:n}^{(2)}, \ldots, \theta_{1:n}^{(B)}$, there is a high probability that the newly sampled θ_{new} from $\hat{G}_0$ is located near the point mass around previous sampled θ. This may result in DP-MBJ overestimating the number of clusters. Also, as the dimension in the dataset increases, the construction of $\hat{G}_0$ using kernel density estimation will not be easy, and will result in overestimating the number of clusters. For other simulation studies, we leave out the step for constructing $\hat{G}_0$ to prevent DP-MBJ from overestimating the number of clusters. Also, because DP-MBJ uses the empirical Bayes approach for estimating the concentration parameter, their estimated concentration parameter is no longer strictly the sample from MCMC chains when $B > 1$. In this section, we conclude that the original DP-MBJ has a tendency to overestimate the number of clusters, and it is more appropriate for density estimation rather than clustering analyses.

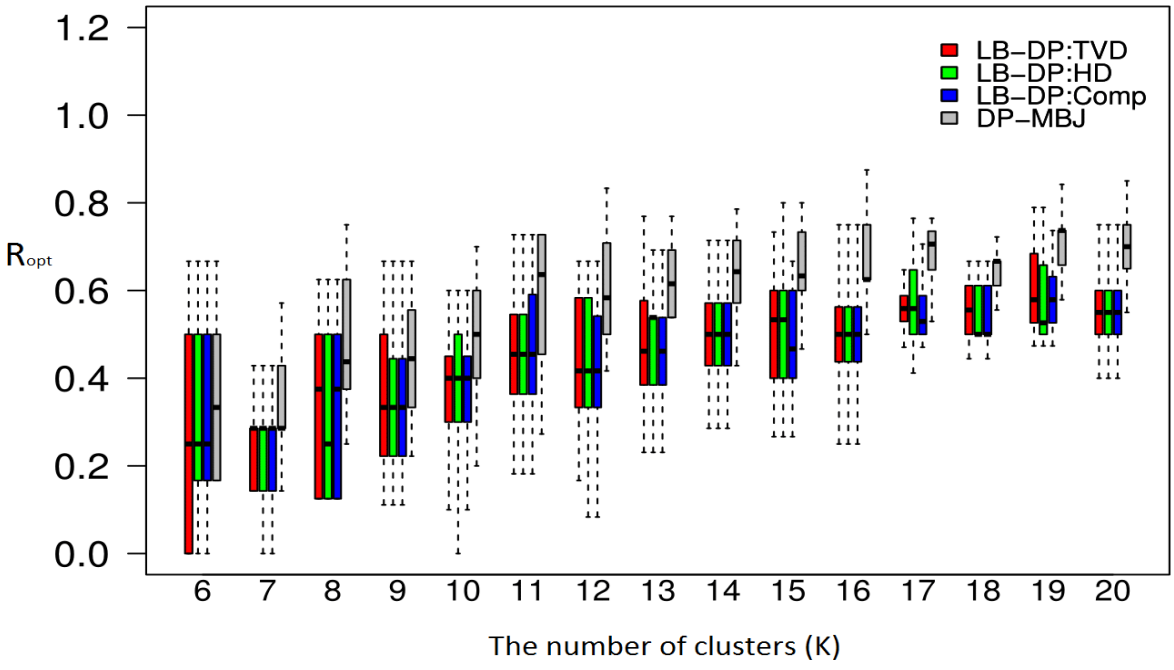


Figure 4.2: Side-by-side boxplots for the distribution of the ratio R_{opt}s obtained by the simulation study. Each boxplot depicts the distribution of the ratio R_{opt} against the true number of clusters K. At different K, the first boxplot from the left (red) describes the distribution of the ratio R_{opt} obtained by LB-DP with total variation distance, and the second boxplot from the left (green) depicts the distribution of the ratio R_{opt} obtained by LB-DP with Hellinger distance. The third boxplot from the left (blue) and the first boxplot from the right (gray) illustrate distributions of the ratio R_{opt} obtained by LB-DP with complete linkage function and by DP-MBJ at each K, respectively.

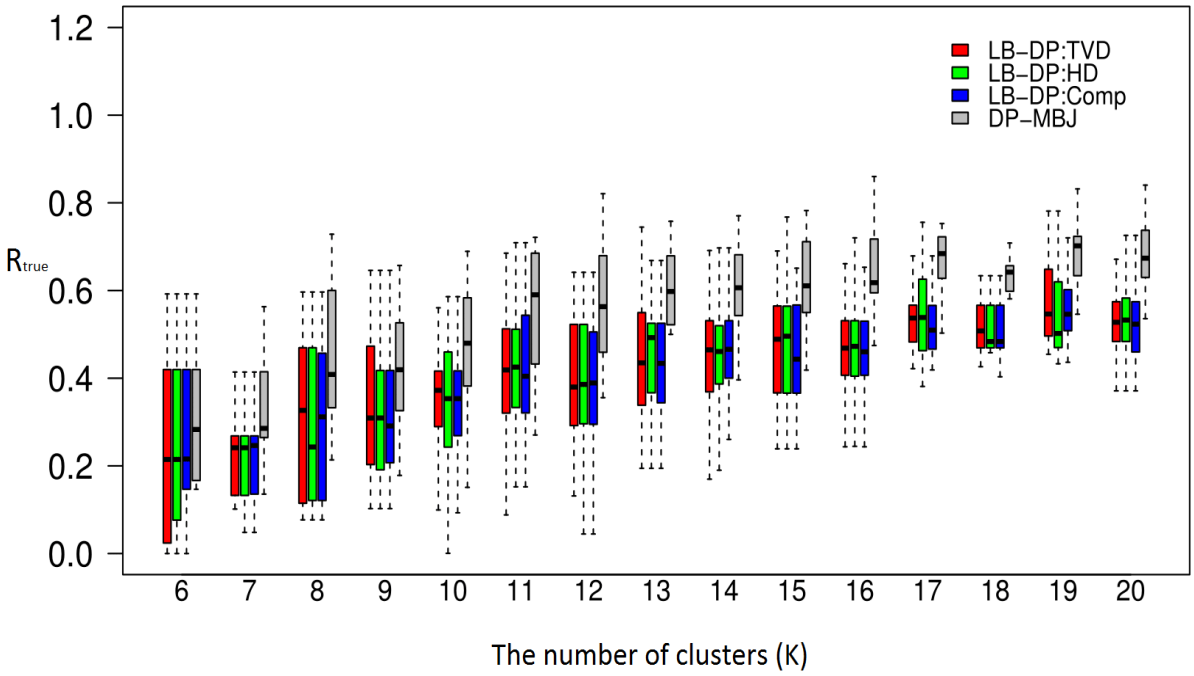


Figure 4.3: Side-by-side boxplots for the distribution of the ratio R_{true}s obtained by the simulation study. Each boxplot depicts the distribution of the ratio R_{true} against the true number of clusters K. At different K, the first boxplot from the left (red) describes the distribution of the ratio R_{true} obtained by LB-DP with total variation distance, and the second boxplot from the left (green) depicts the distribution of the ratio R_{true} obtained by LB-DP with Hellinger distance. The third boxplot from the left (blue) and the first boxplot from the right (gray) illustrate distributions of the ratio R_{true} obtained by LB-DP with complete linkage function and by DP-MBJ at each K, respectively.

4.2.2 Mixture of Bivariate Gaussians with Small Number of Observations

When the number of observations is small, the influence of the concentration parameter becomes considerable. According to the CRP, the probability that n^{th} customer chooses the new table is $\frac{\alpha}{n-1+\alpha}$ in theory. For example, given $n = 1000$ and $\alpha = 5$, this probability becomes $\frac{5}{1004}$. However, given $n = 30$ and $\alpha = 5$, the probability of choosing the new table is $\frac{5}{34}$, which is larger than $\frac{5}{1004}$. In theory, the expected number of clusters from DP in Liu (1996) is exponentially growing when the number of observations is smaller than 50 as shown in Figure 4.4. However, we can see that the expected number of clusters is slowly increasing when the number of observations is greater than 50. In our work, "small" indicates the number of observations is smaller or equal to 50 where the expected number of clusters increases exponentially. In this section, in order to assess the effect of the concentration parameter when we have a low number of observations ($n = 50$), we conduct the simulation studies under two conditions: (1) non-overlapping clusters and (2) overlapping

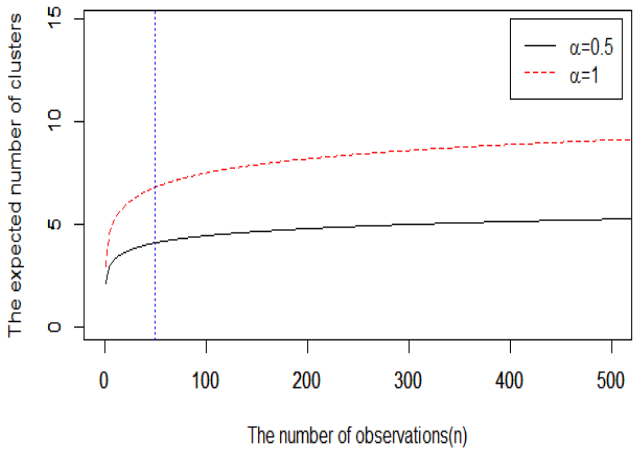


Figure 4.4: The expected numbers of clusters from the DP in theory when $\alpha = 0.5$ and $\alpha = 1$ over the number of observations (n) are plotted. The blue dashed line indicates $n = 50$.

clusters. Under the first condition (non-overlapping clusters), we intend to verify whether linkage based Dirichlet process mixture model works well as a clustering tool for the well-separated clusters. Under the second condition (overlapping clusters), we designate that LB-DP merges highly overlapping clusters into one cluster successfully. We use $K = \{3, 4, 5, 6\}$, and we simulate different sets of data 350 times for each K. We obtain the clustering assignment for each observation based on its clustering assignment probability. Then, we calculate the silhouette coefficient as a metric to measure the performance of clustering models (Rousseeuw, 1987). A silhouette coefficient measures how similar a data point is to its assigned cluster compared to other clusters. The range of the silhouette coefficient is from -1 to 1, and a higher value of the silhouette coefficient indicates a better clustering solution. After data points are clustered, the silhouette $s(i)$ for a data point i is calculated by:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

where $a(i)$ is the average distance of the data point i between all other data points belonging to the same cluster, and $b(i)$ is the lowest average distance of the data point i to other clusters that do not include i. The average of $s(i)$ over the data points in the cluster measures how the data points in the clusters are well-grouped. Then, the average of $s(i)$ over the all data points in the dataset measures the clustering quality how all data points are well-clustered.

Non-overlapping clusters

We demonstrate the simulation result for 50 data points from K non-overlapping mixture of bivariate Gaussian distributions, where $K = \{3, 4, 5, 6\}$. Figure 4.5 displays the distributions of the silhouette coefficients obtained by applying DP-MBJ and LB-DP:HD for non-overlapping clusters. We can see that the distribution of the silhouette coefficients for DP-MBJ (white) and the distribution of the silhouette coefficients for LB-DP:HD (dark-gray) look similar to each other. When we have non-overlapping clusters, LB-DP penalizes the concentration parameter less, since distances between clusters will be close to 1. This leads both methods to have similar silhouette coefficients. We conclude that the two meth-

ods have the similar level of performance with respect to the silhouette coefficients when clusters are not overlapping.

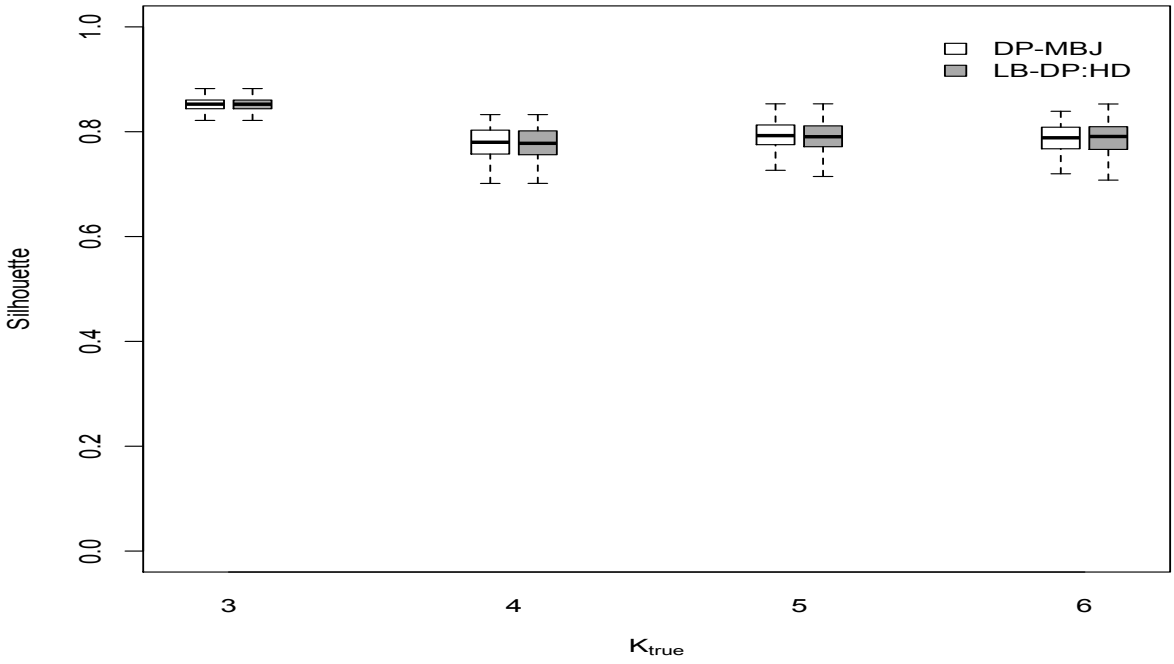


Figure 4.5: Side-by-side boxplots for the distribution of the silhouette coefficient obtained by applying DP-MBJ and LB-DP:HD for non-overlapping clusters in Section 4.2.2.

Overlapping clusters

We analyze the simulation result based on the silhouette coefficients. Figure 4.6 depicts the distributions of the silhouette coefficients obtained by applying DP-MBJ and LB-DP:HD for overlapping clusters. Specifically, as shown in Figure 4.6, the silhouette coefficients obtained by LB-DP:HD have higher medians of the silhouette coefficients than DP-MBJ when we have $K_{true} = 4$ and $K_{true} = 6$. As shown in Figure 4.1(b) and Figure 4.1(d), the number of pairs of overlapping clusters are 2 and 3 for $K_{true} = 4$ and $K_{true} = 6$, respectively. We can infer that the reason LB-DP:HD shows high medians of the silhouette coefficients for $K_{true} = 4$ and $K_{true} = 6$ is because of the fact that the linkage based Dirichlet process

works more powerfully in merging two overlapping clusters into one cluster than DP-MBJ by penalizing the concentration parameter. Especially, when we have a random noise data point, the linkage based Dirichlet process will attempt to merge this data point to the closest cluster, but DP-MBJ leaves this data point as a standalone cluster, and this influences the estimation of the concentration parameter. As an example, Figure 4.7 shows that a data point is claimed to be the one cluster by DP-MBJ. The silhouette coefficients obtained by DP-MBJ and LB-DP:HD are 0.62 and 0.79, respectively for the clustering solutions in Figure 4.7. Linkage based Dirichlet process mixture model performs slightly better than DP-MBJ when we have low number of observations from highly overlapping clusters.

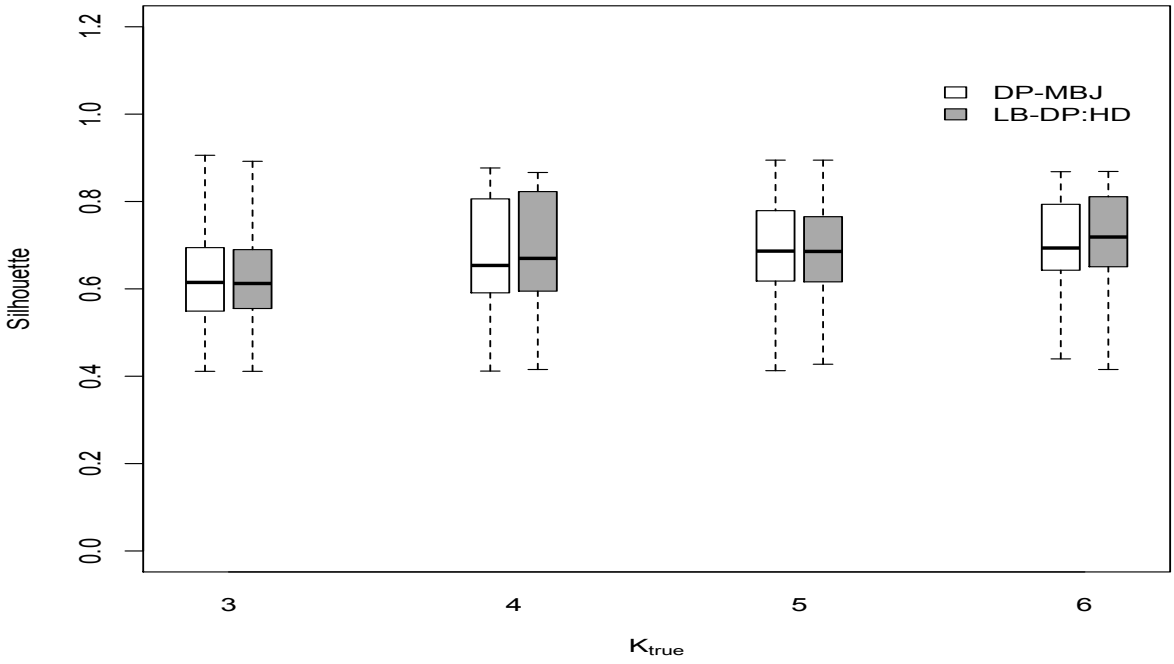
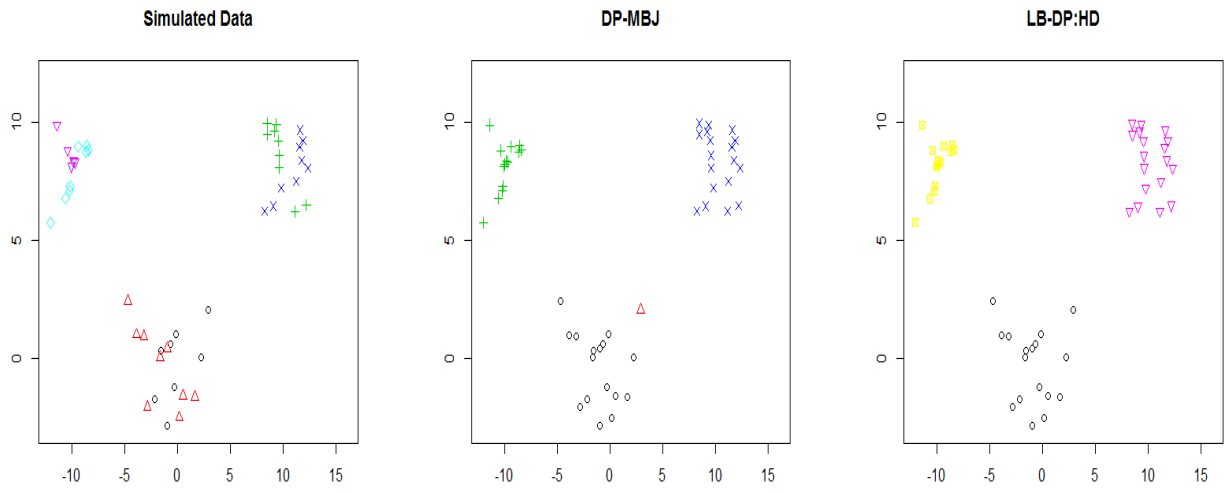


Figure 4.6: Side-by-side boxplots for the distributions of the silhouette coefficients obtained by applying DP-MBJ and LB-DP:HD for overlapping clusters in Section 4.2.2.



(a) True Clusters (b) DP-MBJ (c) LB-NDP

Figure 4.7: Panel (a) depicts the simulated data with their true clustering memberships. Panel (b) shows the clustering solution obtained by DP-MBJ. Panel (c) depicts the clustering solution obtained by LB-DP:HD. The silhouette coefficients for the clustering solutions in Panel (b) and (c) are 0.62 and 0.79, respectively.

4.2.3 Mixture of 5-D Multivariate Gaussians with Small Number of Observations

As an extension of the simulation study in Section 4.2.2, we simulate 50 data points from the mixture of 5-D multivariate Gaussian distributions. We use $K = \{3, 4, 5, 6\}$, and we simulate different sets of data 200 times for each K.

Non-overlapping clusters

As shown in Figure 4.8, DP-MBJ and LB-DP:HD have similar levels of clustering performance with respect to the obtained silhouette coefficients when we have a small number of observations from 5-D dimensional space. The distributions of the silhouette coefficients obtained by DP-MBJ (white) and LB-DP:HD (dark-gray) look similar to each other for each K_{true}. This result is consistent with the one for non-overlapping clusters in Section 4.2.2. We confirm that both have a similar level of performance based on the calculated silhouette coefficients when clusters are very distinct from each other.

Overlapping clusters

As depicted in Figure 4.9, LB-DP:HD has higher medians of the silhouette coefficients than the ones obtained by DP-MBJ. The distributions of the silhouette coefficients obtained by DP-MBJ (white) and LB-DP:HD (dark-gray) look similar to each other for each K_{true}. This result is consistent with the one for overlapping clusters in Section 4.2.2. This implies that LB-DP has the better performance as a clustering tool than DP-MBJ given the small number of data points on 5-D space.

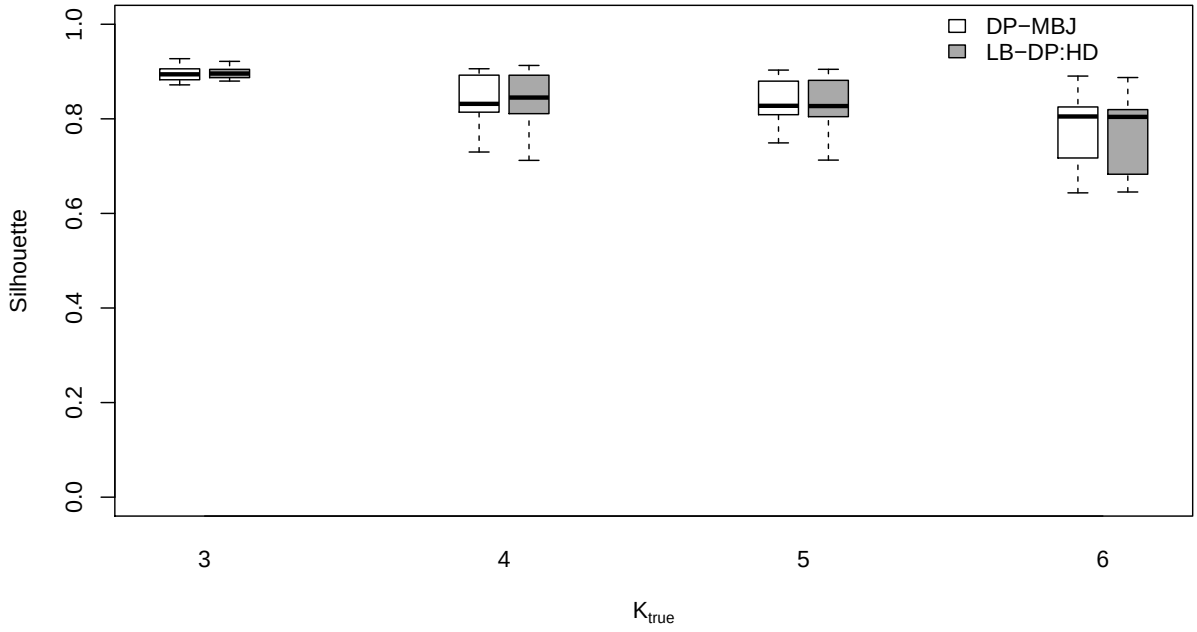


Figure 4.8: Side-by-side boxplots for the distribution of the silhouette coefficient obtained by applying DP-MBJ and LB-DP:HD for non-overlapping clusters on 5-D dimensional space in Section 4.2.3.

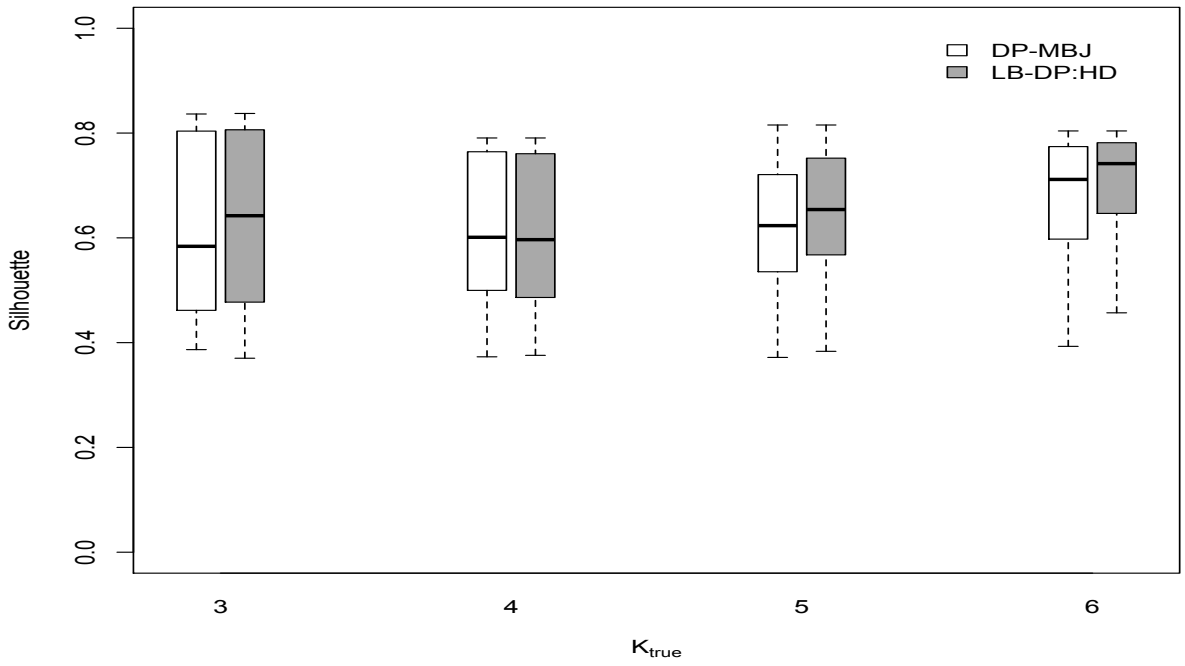


Figure 4.9: Side-by-side boxplots for the distribution of the silhouette coefficient obtained by applying DP-MBJ and LB-DP:HD for overlapping clusters on 5-D dimensional space in Section 4.2.3.

4.2.4 Mixture of 5-D Multivariate Gaussians with Large Number of Observations

In this section, we explore simulation studies for a large number of observations ($n = 1000$) under two different conditions: (1) non-overlapping clusters and (2) overlapping clusters. We focus on examining the influence of the estimated concentration parameter on clustering quality when we have a large number of observations. For performing the linkage based Dirichlet process mixture model, we employ two different distance measures: Hellinger distance and complete linkage function. The reason that Hellinger distance is chosen instead of total variation distance is for easier computation when data dimension $d > 1$. In our simulation, $K = 5, 6, \ldots, 10$. The mixture proportion is drawn from a Dirichlet distribution with K and $\alpha = 1$. For each K, we simulate different sets of data 100 times.

Non-overlapping clusters

Figure 4.10 depicts the distributions of the silhouette coefficients obtained by LB-DP:HD, LB-DP:Comp, and DP-MBJ against $K_{true}(= K)$. We can see that for non-overlapping clusters, both our proposed models, LB-DP:HD and LB-DP:Comp, and DP-MBJ have similar levels of clustering performance. It seems that when $K_{true} > 7$, the medians of the silhouette coefficients obtained by LB-DP:HD are slightly larger than the medians of the silhouette coefficients obtained by DP-MBJ. Also, the choice of probability distance measures in linkage based Dirichlet process prior, either Hellinger distance or scaled complete linkage function, shows that they are not very different with respect to modeling performance.

Overlapping clusters

Figure 4.11 depicts the distributions of the silhouette coefficients obtained by three different models: LB-DP:HD, LB-DP:Comp, and DP-MBJ. Figure 4.11 shows that the modeling performances obtained by the three different methods are similar to each other.

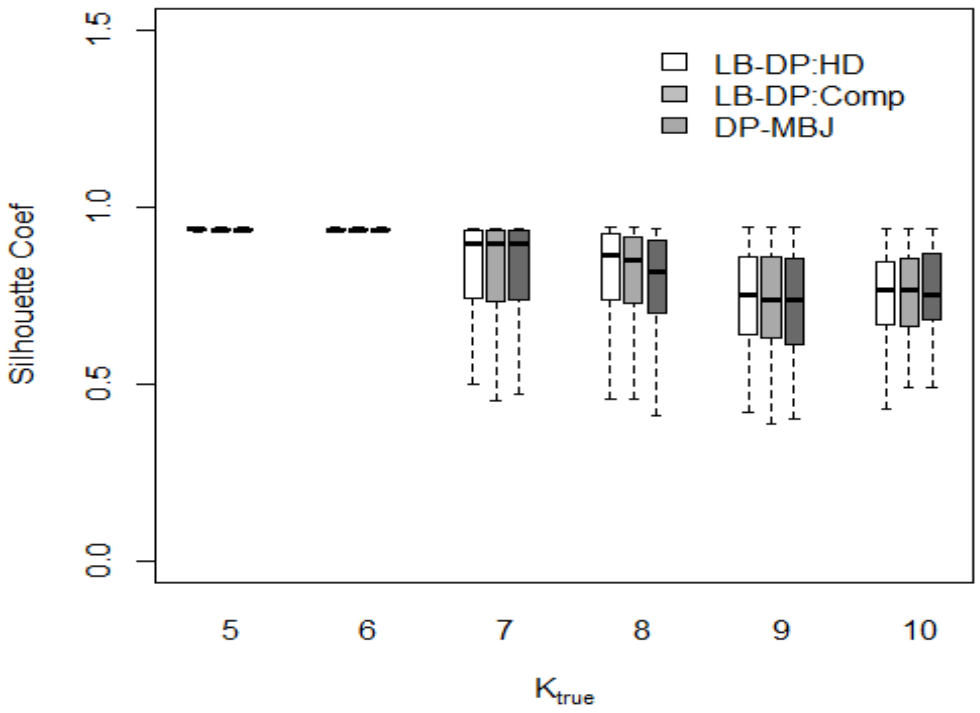


Figure 4.10: Side-by-side boxplots of silhouette coefficients for simulated non-overlapping clusters in Section 4.2.4. At different K_{true}, the left boxplot (white) describes the distribution of silhouette coefficients from applying LB-DP:HD and the boxplot(gray) in the middle depicts the distribution of silhouette coefficients from LB-DP:Comp. The right boxplot (dark gray) describes the distribution of silhouette coefficients from applying DP-MBJ.

Through the simulation study for non-overlapping/overlapping clusters in this section, we have seen that linkage Based Dirichlet process mixture model has similar levels of performance as DP-MBJ. The reasons that we have similar modeling performance from LB-DP and DP-MBJ are possibly the number of observations and their clustering structures. In theory, we know that the number of clusters defined via the Dirichlet process mixture model is dependent on the concentration parameter. However, if we have a large number of observations, the effect of the concentration parameter on having new clusters in the Dirichlet process is minor. Even though, in theory, selecting an appropriate concentration parameter in any type of Dirichlet processes is important to derive clustering solutions, the effect of

the concentration parameter on clustering solution is relatively weak in practice, especially when n is large.

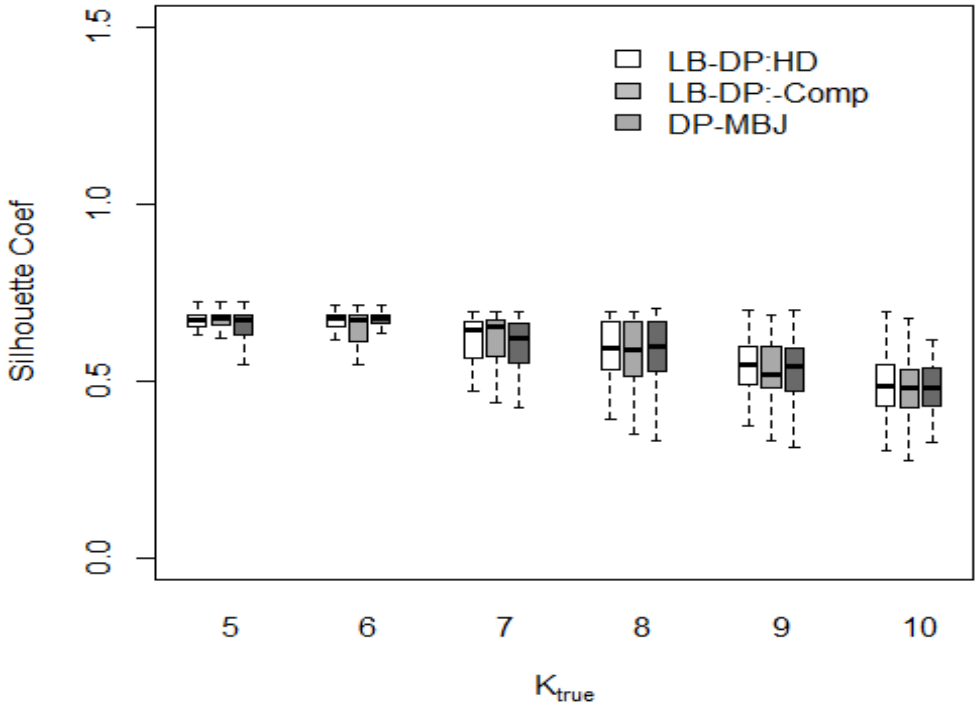


Figure 4.11: Side-by-side boxplots of silhouette coefficients for simulated overlapping clusters in Section 4.2.4. At different K_{true}, the left boxplot (white) describes the distribution of silhouette coefficients from applying LB-DP:HD, the boxplot(gray) in the middle depicts the distribution of silhouette coefficients from LB-DP:Comp while the right one (dark gray) describes the distribution of silhouette coefficient from applying DP-MBJ.

4.3 Conclusions

By conducting a sequential of simulation studies, we have examined whether the linkage based Dirichlet process prior estimates the concentration parameter which specifies the sufficient number of clusters. In Section 4.2.1, given simulated univariate Gaussian mixtures, we investigated the inferred number of clusters. We have demonstrated that the original DP-MBJ that estimates both the concentration parameter and the base measure via kernel density estimation is not appropriate for clustering analysis, because of its inherent tendency of misjudging the number of clusters due to the estimated base measure. In addition, the estimated concentration parameter via MCMC chains no longer possesses the Markov property when $B > 1$. We conclude that the original DP-MBJ is not appropriate for clustering analyses.

A set of simulation studies in Sections 4.2.2, 4.2.3, and 4.2.4 has been performed to verify whether the estimated concentration parameter and the inferred clusters construe the structure of the simulated datasets. We have dropped the step for estimating G_0 for the simulation studies in Sections 4.2.2, 4.2.3, and 4.2.4. Based on the silhouette coefficients, Sections 4.2.2 and 4.2.3 compared the performance of the linkage based Dirichlet process against DP-MBJ for $n = 50$ data points from the mixture of non-overlapping Guassians. It turns out that the linkage based Dirichlet process has a similar level of performance as DP-MBJ given a small number of observations from distinct clusters. However, the linkage based Dirichlet process performs better than DP-MBJ for a small number of observations from highly overlapping clusters. This is because the effect of the size of the estimated concentration becomes relatively large when a small number of data points are observed compared to a large number of data points. Specifically, we conclude that the estimation of the concentration parameter is important for a small number of observed data points.

In Section 4.2.4, we have compared the performance of linkage based Dirichlet process mix-

ture model with DP-MBJ for a large number of observations from overlapping clusters and non-overlapping clusters. What we have found through the simulation studies in Section 4.2.4 is that the concentration parameter has little influence on clustering results defined by Dirichlet process mixture models when the number of observations is substantially large.

Chapter 5

Linkage Based Dirichlet Process: Application

In Chapter 5, we present an application of the linkage based Dirichlet process: modeling the timeline for building construction costs at Virginia Tech. In particular, we apply a mixture model with a linkage based Dirichlet process and a mixture model with a DP-MBJ into the building construction data. Then, we compare the results and infer estimated and classified curves for understanding the pattern of the building construction costs.

5.1 Modeling the Timeline for Building Construction Costs

5.1.1 Background

In functional data analyses, fitting and clustering curves have been prominent applications of the data, where data points correspond to a curve (Lancaster and Salkauskas, 1986; Motulsky and Christopoulos, 2004; Ramsay, 2006; Silverman and Ramsay, 2005; Tarpey, 2007).

Curve fitting is a method of constructing curves or mathematical functions, which approximately fit a series of data points (Kolb, 1983; Pyle, 1999; Motulsky and Christopoulos, 2004). Curve fitting is intimately related to nonlinear regression when data transformation does not work for fitting a linear regression model ((Silverman, 1985; Motulsky and Ransnas, 1987)). By assigning a function to the entire range of data observed with random errors, this method attempts to capture a trend of data points. For example, curve fitting is used to estimate disease progress (Berger, 1981), survival rate (Motulsky and Christopoulos, 2004), and growth rate (Blasco et al., 2003). Blasco et al. (2003) utilizes a Bayesian approach to estimate the growth rate of rabbits in order to compare growth rates from a control group and a treatment group. To fit growth rate curves, Blasco et al. (2003) assumes that the individual rabbit growth rate can be described by the Gompertz function (Gompertz (1825)). Also, Berger (1981) constructs a Gompertz model and a logistic model to capture the trend of plant disease progress. In economics, estimating economic growth rate curve has been an important application in signposting the growth of an economy (Nadaraya, 1964; Hardle, 1990).

With respect to both clustering and estimating curves, Tarpey (2007) introduces the K-means clustering algorithm to segment the functional data by plugging estimated regression coefficients from individual curves. On the other hand, curve fitting can be performed after implementing the K-means algorithm. However, these two approaches are inefficient because we need to either analyze the curves from different clusters or segment estimated individual curves. As discussed in Ray and Turi (1999) and Tibshirani et al. (2001), the K-means clustering method requires attention when choosing the number of clusters for data. Also, the clustering solution from ad-hoc clustering algorithms such as K-means clustering do not provide any probabilistic inference (Fraley and Raftery, 2002). Gaffney and Smyth (2003) proposes random effects regression mixtures, a model-based curve clustering, which enables us to cluster and estimate curves. However, Gaffney and Smyth (2003) also has an issue with choosing the number of clusters. To overcome the issue of selecting the number of clusters, Heard et al. (2006) uses the prior for the number of clusters, which is a uniform prior from

1 to N. Assigning this uniform prior, the size of clusters becomes the multinomial-Dirichlet prior. Heard et al. (2006) utilizes a Bayesian model-based hierarchical clustering algorithm for curve data in order to investigate regulation mechanisms in the genes. However, these models have revealed that some prior information about the number of clusters is needed.

Recently, Dirichlet process mixture models have been used in many areas including nonlinear regression, classification, and density estimation (Susarla and Van Ryzin, 1976; West and Escobar, 1993; MacEachern and Müller, 1998; Escobar and West, 1995; Müller et al., 1996). A Dirichlet process prior in the model, which exhibits the clustering effects, enables us to perform simultaneous clustering and density estimation. The major advantage of using the Dirichlet process mixture model is that this model is free from pre-assigning the number of clusters (Figueiredo and Jain, 2002; Dahl, 2006). For example, the Dirichlet mixture model has its use for estimating and grouping curves without assigning the number of clusters (Müller et al., 1996; Shahbaba and Neal, 2009; Nsoesie et al., 2014). However, the clustering solution is dependent on the selection of the concentration parameter. In Chapter 5, we utilize the mixture model with a linkage based Dirichlet process which helps in both gaining the appropriate number of clusters and estimating the reasonable size of the concentration parameter for simultaneous curve estimation and clustering.

5.1.2 Data

As academic universities are required to be answerable for all properties under their control by federal laws, the office of Capital Assets and Financial Management in Virginia Tech coordinates long-term strategic plans for its capital improvements and financial managements. This includes budget development and financing, and forms the basis for authorizing major capital projects. Also, this office invests in a broad range of capital assets which include land and land improvements, building and building improvements, facilities and other im-

provements, etc. In particular, constructing new buildings contributes to an increase in the value of a university's capital assets, and building improvements extend the life of an existing building. When there are construction projects, the Office of Capital Assets and Financial Management draws up expense budgets for long-term constructions such as constructing new buildings and for short-term constructions such as remodeling ramshackle facilities. Then, the office of Capital Assets and Financial Management organizes the processes of projects and supports the delivery of accountable, competitive and diverse resources for those projects smoothly. This office is needed to avoid any penalties caused by failing to deliver the resource on time, such as late payments. Thus, when managing budgets and purchasing facilities for the university, it is very important to estimate future expenses in advance so that the office can disburse money at the proper time. This chapter uses data that focus on occurrences of expenses in building construction projects at Virginia Tech. We aim not only to cluster building construction projects based on their cumulative expenditure rates, but also to estimate clustered curves for describing features of cumulative expenditure rate curves.

Our data contains 30 cumulative expenditure rates corresponding to building construction projects at Virginia Polytechnic Institute and State University from 2004 to 2011. In general, the construction project has 3 phases: a design, a construction, and a closeout phase (Gould and Joyce, 2003; Clough et al., 2000; Fisk, 1988). We can view the design phase as a refinement of the scope of the project after identifying and reviewing the construction project proposal. The design phase outlines, coordinates, and confirms the scope of the project and detailed plans for the elements of the project such that equipment installations, landscape, and budget (Clough et al., 2000; Fisk, 1988; Trauner, 1993). Then, the construction phase is the actual period when contractors start building. At the construction phase, the budget will be spent on delivering facilities, installation, and real constructions (Clough et al., 2000; Fisk, 1988). In general, the construction phase eats up the major proportion of the project budget. As the final stage of the project, the closeout phase is the period for facilitating, coordinating, and organizing occupancy (Fisk, 1988; Trauner, 1993; Clough et al., 2000).

Thus, compared to other phases, a small portion of the budget goes into the closeout phase. As an example, the cumulative rate of the construction project through construction phases is depicted in Figure 5.1. The cumulative expenditure rate in the design phase grows linearly, and the one in the closeout phase is nearly flat. However, the cumulative rate in the construction phase grows rapidly. Also, a high proportion of the project budget is spent in the construction phase. Thus, it seems that there are change points with respect to the cumulative expenditure rate between phases. Figure 5.1(a) shows the expenditure rate curve for the building project, "Chemistry/Physics-Phase II", over the construction phases. Figure 5.1(a) depicts that the expenditure rate in the design phase grows linearly and the one in the closeout phase is nearly flat. However, the cumulative cost rate in the construction phase grows exponentially and a high proportion of the project budget is spent in the construction phase. Thus, it seems that there are trade-offs with respect to the cumulative expenditure rate between phases. The building construction projects have different construction periods in months. Among 30 projects, the shortest construction period is 19 months. Thus, for analysis we are required to scale the length of the construction period for all projects to 19. Thus, the cumulative cost rates have been modified in accordance with the scaled construction periods by calculating the weighted averages. We provide how the weighted averages are calculated to scale the construction data:

Step 1. For building construction project j, denote the original length of data and data points be T and x_t respectively, where $t = \{1, 2, ..., T\}$.

Step 2. For $i \in \{1, 2, ..., 19\}$, find two data points, $x_{[\frac{T}{19}]i}$ and $x_{[\frac{T}{19}]i+1}$.

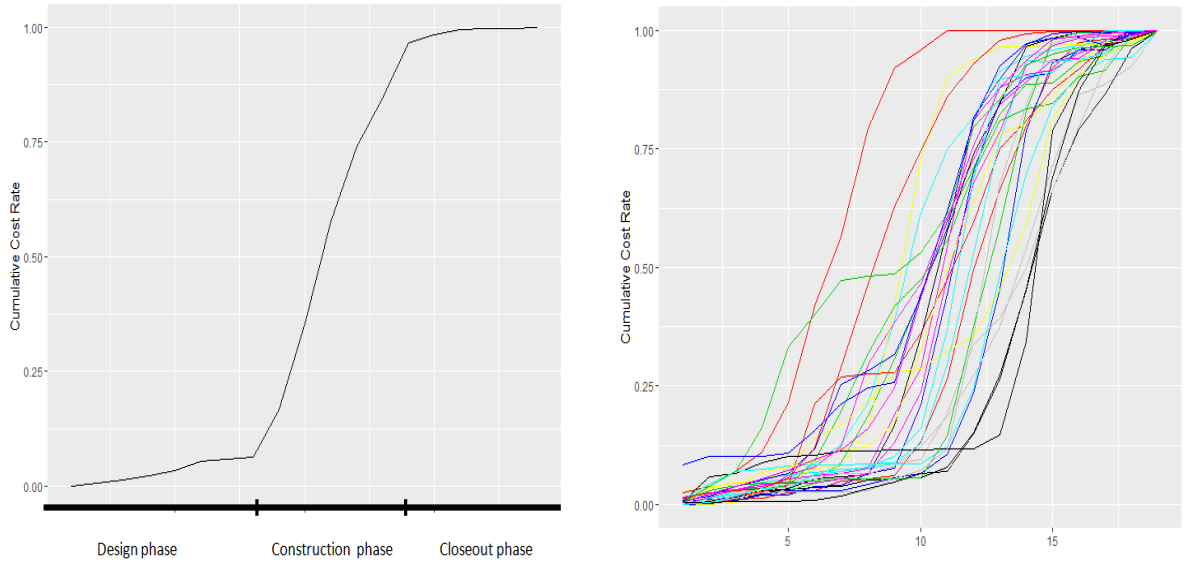
Step 3. Solve the following equation with respect to $x_{new,i}$:

$$x_{new,i} = \frac{x_{[\frac{T}{19}]i}\frac{1}{d_L} + x_{[\frac{T}{19}i+1]}\frac{1}{d_R}}{\frac{1}{d_L} + \frac{1}{d_R}},$$

where d_L and d_R are the euclidean distances between $x_{new,i}$ and $x_{[\frac{T}{19}]i}$ and between $x_{new,i}$ and $x_{[\frac{T}{19}]i+1}$.

Figure 5.1(b) visualizes 30 scaled cumulative cost rate curves corresponding to 30 construction projects at Virginia Tech and we can see that some of curves have distinct patterns.

We can apply the K-means clustering algorithm in order to classify the building projects. When using the K-means clustering algorithm, the optimal number of clusters is not determined in a statistical way. We may rely on the scree plot in Figure 5.2 to choose the number of clusters. However, as shown in Figure 5.2, the within groups sum of squares is gradually decreasing over the number of clusters larger than 6, and this is not informative when choosing the optimal number of clusters. Thus, we decide to use a linkage based Dirichlet process mixture model that automatically determines the optimal number of clusters by updating the concentration parameter and estimates the curve simultaneously.



(a) (b)

Figure 5.1: Panel (a) describes the pattern of the cumulative cost rate for the building construction project , "Chemistry/Physics- Phase II", over design phase, construction phase, and closeout phase. Panel (b) depicts 30 scaled cumulative cost rate curves corresponding to 30 building construction projects at Virginia Tech.

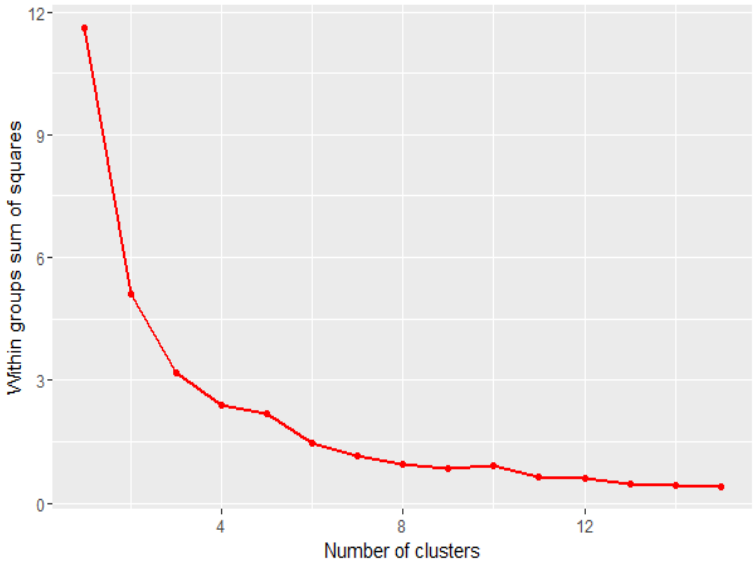


Figure 5.2: The scree plot depicts the Within groups sum of squares according to the number of clusters in K-means clustering analysis for building projects.

5.1.3 Model Specification

We explore a parameterization of the basis function for modeling the timeline for building construction costs. Then, we form a linkage based Dirichlet process mixture model with Hellinger distance and a mixture model with DP-MBJ in order to partition 30 curves into K clusters, where the range of possible K is between 1 and 30.

Given time $t = \{1, 2, \ldots, 19\}$, $i = \{1, 2, \ldots, 30\}$, and $y_i = \{y_{i,1}, y_{i,2}, \ldots, y_{i,19}\}$ corresponds to the curve of the building construction project i, the model is as follows:

$$y_{it} = f(\theta, t) + \epsilon_{it},$$
$$\epsilon_{it} \sim N(0, \sigma_i^2),$$

where $f(\theta, t)$ is the function of time t and the parameter θ, and error terms ϵ_{it} are independent and identically distributed. As a basis function $f(\theta, t)$, we may consider the use of the Gompertz function (Gompertz, 1825). In the following, consider the described Gompertz

function:

$$f(t) = ae^{-be^{-ct}}, \tag{5.1}$$

where $t \in (-\infty, \infty)$, a is an asymptotes, and positive values b and c are shape parameter and scale parameter, respectively. The Gompertz function is the favorable basis function in order to model the growth rate in biology (Berger, 1981). However, in this work, it is necessary to modify the Gompertz function by adding more parameters for following reasons: 1) considering the shape of our cost rate curves; it seems the curve is a mixture of linear and Gompertz functions; 2) We need to explain where changes from a linear growth to an exponential growth or an exponential growth to a linear growth have taken place. In particular, the change-point, where the curve grows exponentially, will be the starting time point where the Office of Capital Assets and Financial Management pays out a high proportion of the budget.

To construct the modified Gompertz function (Clarke et al., 2013), first, we set the parameter b in Equation 5.1 to be 1 for all curves as this parameter explains the displacement of the curves from the left to the right, which should be same for all curves in our data. Also, in terms of t, we add a shift parameter since our data has $t \geq 0$ and this shift parameter explains where the curve grows exponentially. To demonstrate the linear pattern in the rate curve, we introduce a slope parameter. Then, we include the offset parameter for describing where the curves are linearly growing after their exponential growth. Finally, our basis function with five parameters is as follows:

$$f(\theta, t) = \theta_1 e^{-e^{-\theta_3(t-\theta_2)}} + I(t < \theta_4) \times \theta_5(t-1) + I(t \geq \theta_4) \times \theta_5(\theta_4 - 1), \tag{5.2}$$

where I is an indicator function. The parameters in Equation 5.2 have different roles. θ_1, θ_2, θ_3, θ_4, and θ_5 represent a height parameter, a shift parameter, a growth rate parameter,

an offset parameter, and a slope parameter, respectively. In particular,

$$\begin{aligned}\lim_{t\to+\infty} f(\theta,t) &= \theta_1 e^{-e^{\infty}} + \theta_5(\theta_4 - 1) \\ &= \theta_1 e^{0} + \theta_5(\theta_4 - 1) \\ &= \theta_1 + \theta_5(\theta_4 - 1).\end{aligned} \tag{5.3}$$

Equation 5.3 shows that θ_1, θ_4, and θ_5 jointly determine an asymptote in our modified Gompertz function. The Gompertz curves and the modified Gompertz curves with different parameter settings are depicted in Figure 5.3. The Gompertz curves (dotted, dot-dashed, and long dashed) in Figure 5.3 show asymptotes are determined by the parameter a. For example, depending on the choice of $a = 0.9$ or $a = 1$, the asymptotes in the curve are different. The modified Gompertz curve (black and solid) and the Gompertz curve (red and long-dashed) look similar in Figure 5.3. However, the modified Gompertz function provides more flexibility in terms of the curve shape than the Gompertz function, especially when the curve grows linearly or when the curve has a change point-where the curve grows exponentially. We employ the modified Gompertz function in Equation 5.2 as our basis function to demonstrate the pattern of the cumulative cost curves in Figure 5.1(b).

Given by the defined basis function in Equation 5.2, our linkage based Dirichlet process mixture model for the cumulative cost rate curves is hierarchically formed as:

$$\begin{aligned} y_i &\sim MVN(f(\theta_i,t), \sigma_i^2 I_{t\times t}), \\ \theta_i, 1/\sigma_i^2 &\sim G, \\ G &\sim LBDP(\alpha, G_0), \end{aligned} \tag{5.4}$$

where $\theta_i = (\theta_{i1}, \theta_{i2}, \theta_{i3}, \theta_{i4}, \theta_{i5})$. For the base measure G_0, we use uniform priors for the parameters in Equation 5.2 because we have information that the reasonable range of parameters and the Gamma$(1,1)$ prior for $1/\sigma^2$. As a linkage function in the linkage based Dirichlet process prior, we utilize Hellinger distance. For DP-MBJ, we use the following

model:

$$\begin{aligned} y_i &\sim MVN(f(\theta_i, t), \sigma_i^2 I_{t\times t}), \\ \theta_i, 1/\sigma_i^2 &\sim G, \\ G &\sim DP(\alpha, G_0), \end{aligned} \tag{5.5}$$

with $B = 100$ samples. We use the same base measure G_0 as the linkage based Dirichlet process mixture model in Equation 5.4.

Curve Simulation

Before we analyze our building construction data, we performed a simulation study to examine whether the linkage based Dirichlet process mixture model successfully classifies the

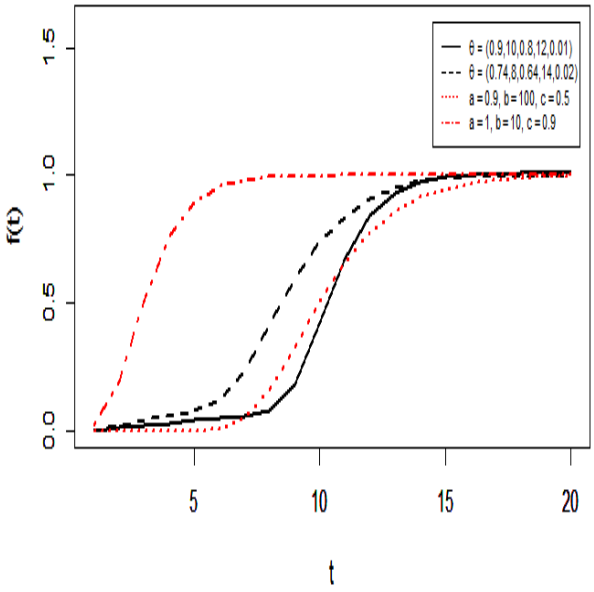


Figure 5.3: Visualization of the Gompertz function and the modified Gompertz function with different parameter settings. The Gompertz function with 2 different parameter settings (dotted and dot-dashed) and the modified Gompertz function with 2 different parameter settings (solid and dashed) are depicted.

groups of curves and estimates the parameters in 5.2. We used five parameter sets in Table 5.1 to simulate 45 curves and $t = \{1, 2, \ldots, 30\}$. 45 simulated curves are depicted in Figure 5.4(a). The model is as follows:

$$\begin{aligned} y_i &\sim MVN(f(\theta_i, t), \sigma_i^2 I_{t\times t}), \\ \theta_i, 1/\sigma_i^2 &\sim G, \\ G &\sim LBDP(\alpha, G_0), \end{aligned} \tag{5.6}$$

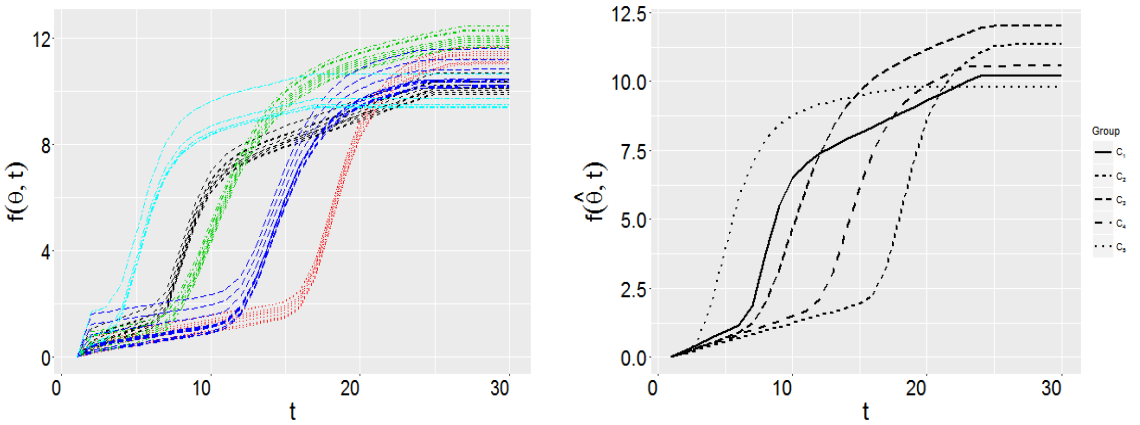
where $i = \{1, 2, \ldots, 45\}$ and $t = \{1, 2, \ldots, 30\}$. Hellinger distance has been used to measure the distance between the clusters. As out base measure, we used the uniform priors for $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5))$ and the $Gamma(1, 1)$ prior for $1/\sigma^2$.

As a result, the linkage based Dirichlet process mixture model based on our basis function

Table 5.1: The true parameters used for generating 45 curves.

Group	# of curves (n)	θ_1	θ_2	θ_3	θ_4	θ_5
1	10	5	8	1	24	0.2
2	10	8	18	0.65	25	0.14
3	10	8	10	0.45	24	0.18
4	10	7	14	0.55	23	0.16
5	5	8.5	5	0.63	18	0.09

successfully identified the original clusters, and estimated the parameters. We provide the estimated parameters in Table with their 95% credible intervals, and depict the posterior distributions of the estimated parameters in Figure 5.5. We also provide five estimated curves by the linkage based Dirichlet process mixture model in Figure 5.4(b). Through the simulation study, we have verified that the linkage based Dirichlet process mixture model successfully classify the curves and estimate the parameters. We also performed DP-MBJ for the simulated curves, and DP-MBJ also found the same clusters. In Section 5.1.4, we present the modeling results for building construction data.

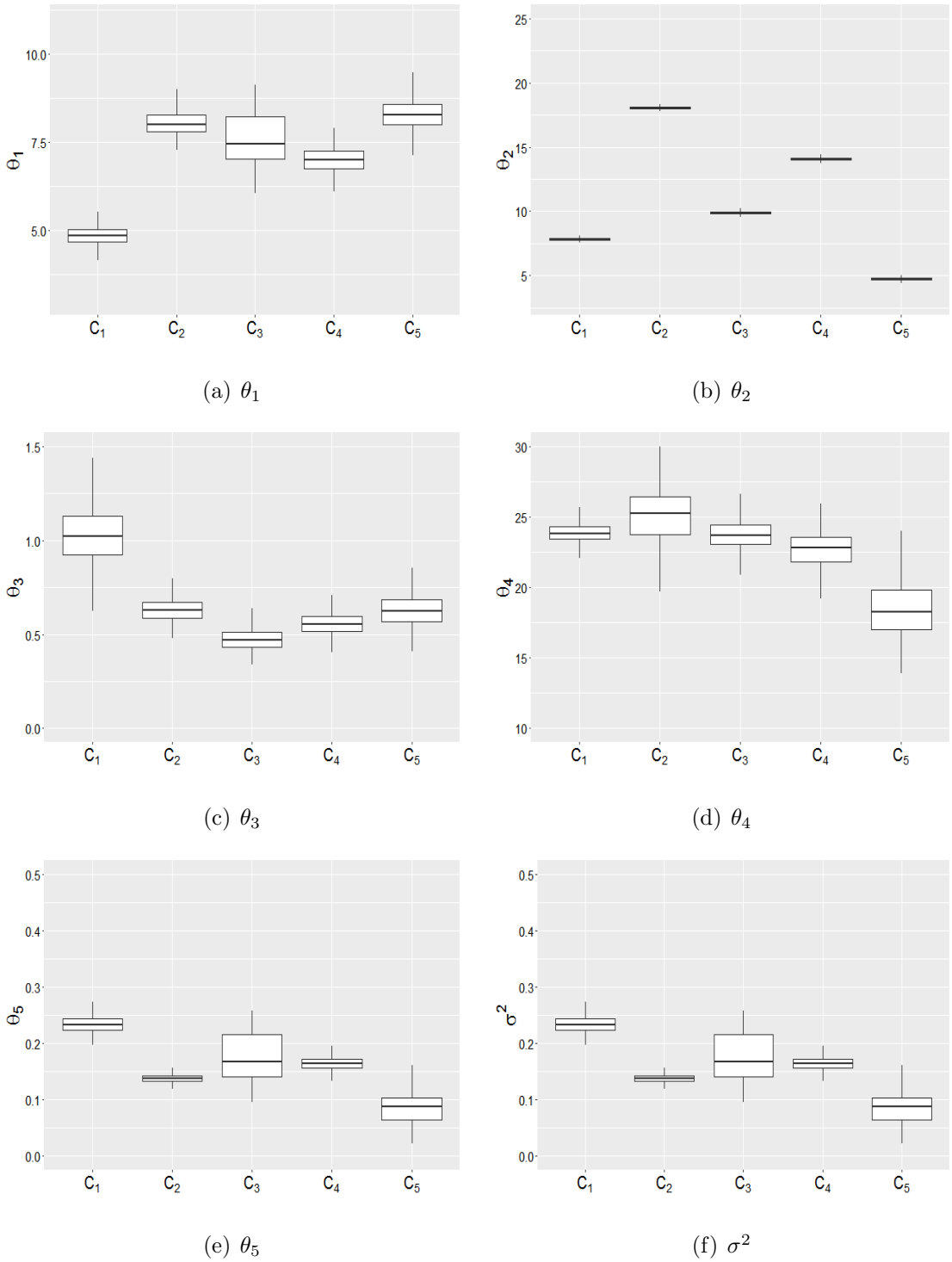


(a) The 45 Simulated Curves (b) The estimated curves

Figure 5.4: Panels (a) and (b) depict the 45 simulated curves and the five estimated curves by the linkage based Dirichlet process mixture model, respectively.

Table 5.2: The estimated parameters by the linkage based Dirichlet process mixture model for five clusters defined in the simulation are summarized.

Group	θ_1	θ_2	θ_3	θ_4	θ_5	σ^2
C_1	4.865	7.827	1.049	23.900	0.233	0.224
	(4.3578,5.3789)	(7.6201,8.0272)	(0.7773,1.4728)	(22.4716,25.3992)	(0.2019,0.2645)	(0.1764,0.3050)
C_2	8.055	18.089	0.631	24.964	0.138	0.170
	(7.5469,8.7737)	(17.8811,18.3437)	(0.5215,0.7528)	(19.4520,28.5991)	(0.1251,0.1509)	(0.1251,0.2466)
C_3	7.895	9.863	0.457	24.369	0.176	0.181
	(6.3752,9.3258)	(9.6308,10.1055)	(0.3640,0.5892)	(22.1434,30.5653)	(0.1071,0.2504)	(0.1259,0.2736)
C_4	7.011	14.074	0.5558	22.694	0.164	0.201
	(6.4126,7.6872)	(13.8397,14.3027)	(0.4591,0.6725)	(20.2316,24.8497)	(0.1416,0.1870)	(0.1527,0.2762)
C_5	8.283	4.716	0.633	18.437	0.086	0.0180
	(7.0626,9.1708)	(4.4847,4.9442)	(0.4972,0.8450)	(15.2774,21.9404)	(0.0298,0.1747)	(0.1342,0.2556)



(a) θ_1 (b) θ_2

(c) θ_3 (d) θ_4

(e) θ_5 (f) σ^2

Figure 5.5: The distributions of the MCMC samples for parameters in Equation 5.2 when LB-DP is applied to simulated curves. Panels (a), (b), (c), (d), (e), and (f) depict the posterior distributions of θ_1, θ_2, θ_3, θ_4, θ_5, and σ^2 by cluster label, respectively.

5.1.4 Results

In this section, we identify the cluster assignments for 30 curves by calculating the cluster assignment probability and demonstrate the timeline for building construction costs. By defining the posterior mode of the number of clusters, we obtain the *optimal* number of clusters along with an estimated concentration parameter.

Sensitivity of the choice of the concentration parameter

Before applying the linkage based Dirichlet process mixture model (LB-DP:HD) and DP-MBJ, we perform the Dirichlet process mixture model for our building project data. This model is used to assess the sensitivity of the number of clusters related to the size of the concentration parameter. We conduct experiments with the various values of the concentration parameter, $\alpha \in \{1, 5, 10\}$. Figure 5.6 compares the estimated number of clusters using the Dirichlet process mixture model with the concentration parameter fixed at $\{1, 5, 10\}$. Figure 5.6 depicts samples of size 10000 for the number of clusters after burn-in for $\alpha \in \{1, 5, 10\}$. As shown in Figure 5.6, note that the posterior distribution for the number of clusters depends on the choice of the concentration parameter (α). As α increases, the number of clusters increases. Thus, in order to gain meaningful groups given the building project data, the estimation of the concentration parameter is essential.

Clustering results

In what follows we compare the clustering results from LB-DP:HD with DP-MBJ. Then, we concentrate on an interpretation of clustering results. As both LB-DP:HD and DP-MBJ provide the various solutions, we investigate our clustering results as follows:

1. After the MCMC chains for parameters in Equation 5.2 converge to the stationary distribution, we calculate the cluster assignment probabilities for curves and decide which cluster a curve belongs to.

2. We obtain the number of clusters from the posterior mode from the MCMC chain.

Then, we examine whether this result is consistent with the findings that we draw by calculating clustering assignment probabilities.

3. We summarize the estimated parameters in our basis function for defined clusters and draw an inference about the concentration parameter in the linkage based Dirichlet process.

Any type of Dirichlet process mixture models suggest various clustering solutions. There is no best way to demonstrate the clustering memberships. However, we can calculate the clustering assignment probability for each building project and determine clustering memberships. We use 10000 samples for class indicator variables, which are drawn from the MCMC chain after a burn-in period for clustering assignment probability calculation. For example, LB-DP:HD concludes that the building project, "Agriculture/Natural Resources Lab", belongs to cluster C_3 in Table 5.4 with the clustering assignment probability 0.9892. For LB-DP:HD, we have identified eight clusters after calculating clustering assignment probabilities for all 30 curves. Also, Figure 5.7(a) shows the posterior mode of the number of

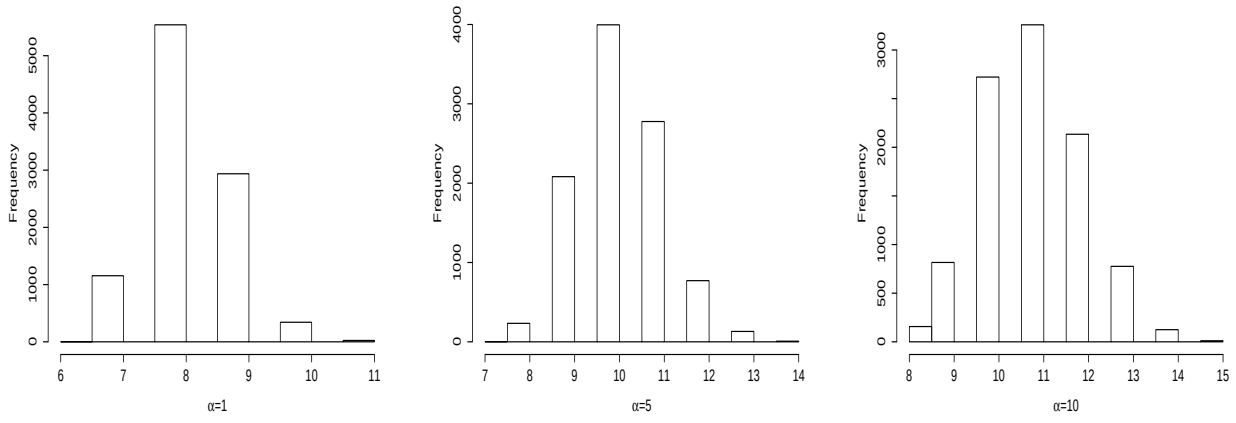
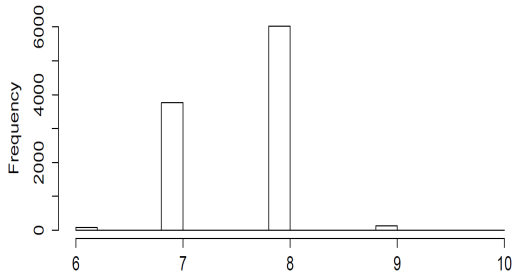
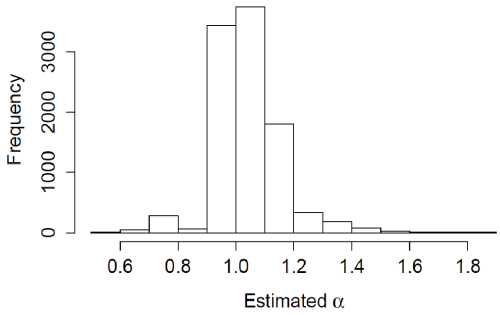


Figure 5.6: A comparison of the estimated number of clusters depending on the choice of the concentration parameter. The posterior distributions of the number of clusters for the building project data defined through the Dirichlet process mixture model with the different size of $\alpha \in \{1, 5, 10\}$ are depicted.

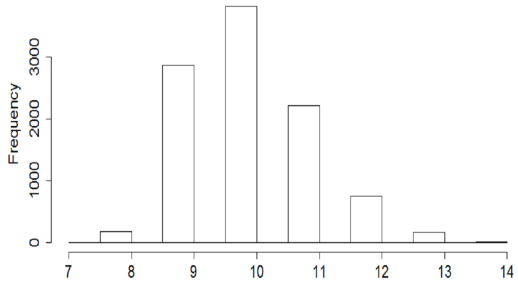
clusters is eight. We conclude that the appropriate number of clusters defined by LB-DP:HD is eight. We describe the posterior distribution α in Figure 5.7(b), and the posterior mean for the concentration parameter is 1.039. The clustering assignment for each building project is in Table 5.4.



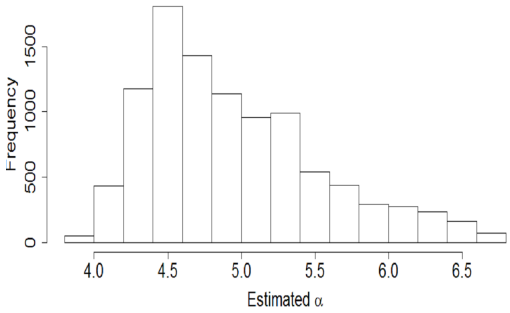
(a) The number of clusters by LB-DP:HD



(b) α by LB-DP:HD



(c) The number of clusters by DP-MBJ



(d) α by DP-MBJ

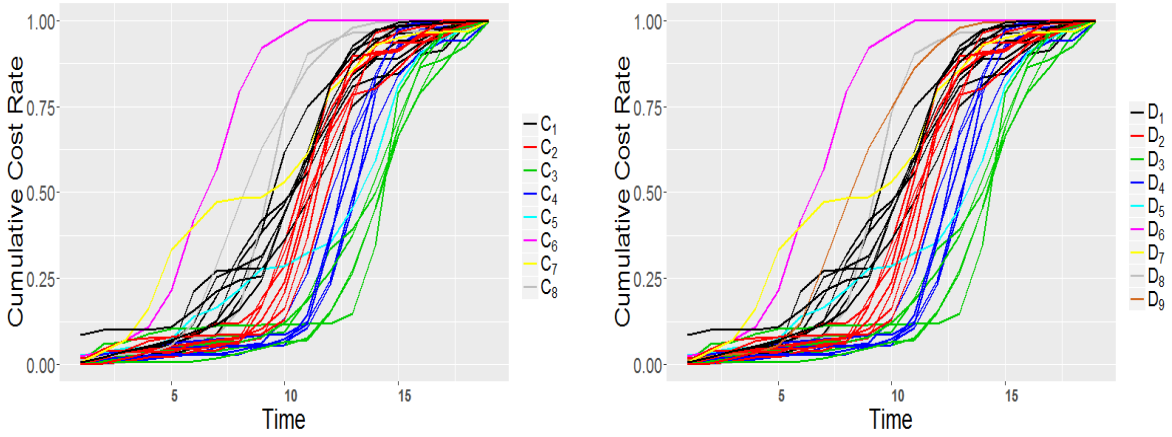
Figure 5.7: Panels (a) and (c) show the histogram of MCMC samples for the number of clusters by applying LB-DP:HD and the histogram of MCMC samples for the number of cluster by applying DP-MBJ, respectively. Panels (b) and (d) are the histograms of MCMC samples for the concentration parameter estimated by LB-DP:HD and estimated by DP-MBJ.

We have identified the number of clusters by the posterior mode of the number of clusters and the number of clusters by calculating the clustering assignment probability for each building project. When LB-DP:HD is applied, these two numbers are same but we have a different result when we use DP-MBJ. With the application of DP-MBJ, we have identified nine clusters after calculating clustering assignment probabilities for all 30 curves. However, Figure 5.7(c) shows the posterior distribution of the number of clusters defined by DP-MBJ, and the posterior mode for the number of clusters is 10. This number is not consistent with the number of clusters based on calculating the clustering assignment probabilities. It may be because the choice of B in DP-MBJ causes mis-estimation of the concentration parameter, which results in the convergence of MCMC chain related to the defined number of clusters. Thus, we may think that the size of the concentration parameter estimated by DP-MBJ leads to the additional standalone cluster at each iteration of the MCMC chain; especially, the posterior mean of sampled α in Figure 5.7(d) is 4.94, which is relatively larger than the one in Figure 5.7(b). This may result in producing more number of clusters than the sufficient number of clusters.

Two methods, LB-DP:HD and DP-MBJ, have brought about different clustering assignment results for two building projects, "Career Services Facility" and "New Residence Hall". Our 30 cumulative cost rate curves are depicted and colored according to the clustering solutions obtained by LB-DP and DP-MBJ in Figure 5.8. When LB-DP:HD is applied, these two building projects belong to the same cluster, which is labeled with C_8 in Table 5.3 and Table 5.4. This result may have come about because LB-DP penalizes the concentration parameter to merge these two building projects into the one cluster, as the probability distance between theses two curves are close to each other. The samples drawn from the MCMC chain for parameters by the cluster label are visualized in Figure 5.9. The clusters labeled with C_1, C_2,..., and C_8 on x-axis in Figure 5.9 are equal to the cluster labels in Table 5.3 and Table 5.4. However, DP-MBJ has defined nine clusters in total when calculating the clustering assignment probabilities, and DP-MBJ has separated these two building projects to be in

the different clusters. We can see that these two building projects labeled with D_8 and D_9 on x-axis in Figure 5.9 have different estimates for θ_2, θ_3, and θ_4, but similar estimates for θ_1 and θ_5. Also, we have compared the silhouette coefficients for clustering solutions obtained by DP-MBJ and LB-DP, and LB-DP has the higher silhouette coefficient (0.32) than the one obtained by DP-MBJ (0.29). Thus, for summarizing parameter estimates and clustering results, we accept the clustering solution defined by LB-DP:HD.

The estimated parameters $\theta = \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5\}$ with their 95% credible intervals in the basis function for eight clusters are given in Table 5.3. As discussed previously, θ_1, θ_4, and θ_5 jointly describe the asymptotes of the expenditure cost curve. θ_4 determines the time when the expenditure rate starts growing slowly and indicates when a closing phase starts. θ_5 illustrates a linear slope in the basis function. Noticeably, estimated parameters for "Football Fields" in group C_6 show that this building project has a distinct pattern of the cumulative expenditure rate in their closing phase. The posterior mean of θ_4 for "Football Fields" is 9.181. This result indicates that this building project has relatively longer periods of the closing phase than building projects in other clusters. This information may be helpful for



(a) LB-DP:HD (b) DP-MBJ

Figure 5.8: Panels (a) and (b) visualize 30 cumulative cost rate curves with colors based on clustering memberships in Tables 5.4 and 5.5, respectively.

managing budget. In addition, building projects in C_3 in Table 5.4 exhibit the shortest period of the closeout phase. As shown in Equation 5.2, our basis function is the mixture of functions, the function of exponential growth and the linear function. θ_1 contributes to determining the height of expenditure cumulative rate curves in the part of exponential growth function. Thus, the estimated θ_1 is the approximate proportion of the expenditure before the end of construction phase. Additionally, relatively large estimated θ_2 in Table 5.3 points to the longer period of the design phase and indicates when a high proportion of budget will start paying out. Therefore, building projects in group C_3, C_4, and C_5 are the construction projects that may require more time on design. Also, this information will be helpful for setting up the timeline for building construction cost. The growth parameter θ_3 is the rate of the exponential growth of expenditure rate curve. Thus, we may conclude that smaller estimated θ_3 has less rapid exponential growth rate. In other words, for example, from the building project, "Football fields", we can observe that the cumulative expenditure rate is increasing at a fast rate. The estimated cumulative expenditure rate curves, which represent groups, are depicted in Figure 5.10. In addition, the estimated cumulative expenditure rate curves that are estimated by DP-MBJ are illustrated in Figure A.1 in Appendix A.

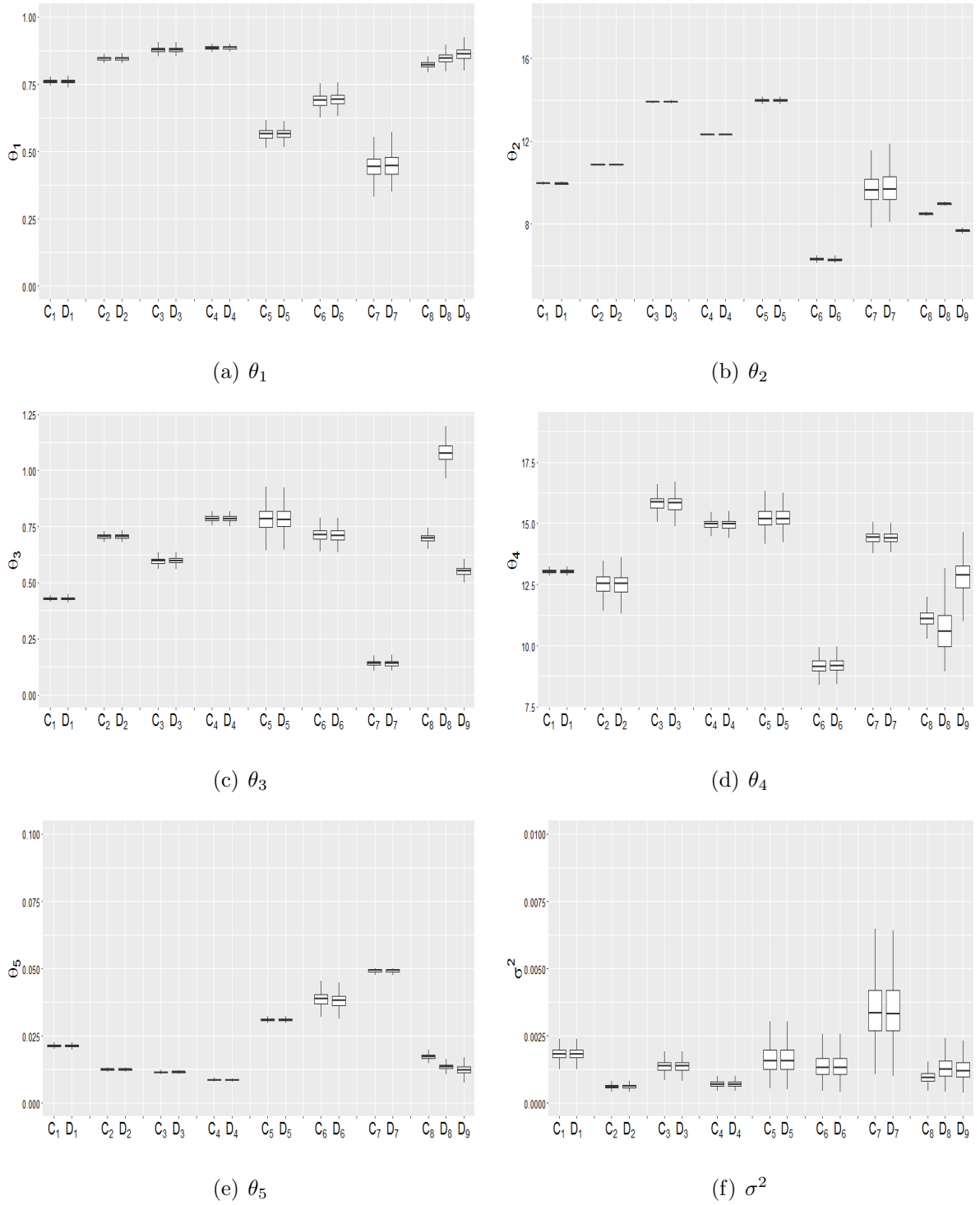
For modeling cumulative expenditure rate curves, we have utilized the linkage based Dirichlet process mixture model and have shown that the linkage based Dirichlet process mixture model have an ability to jointly estimate curves and define clusters. Also, in our application, a part of our curves has a form similar to that of the Gompertz growth curves, which means that we can apply our model to any application such as estimating growth rates. It would be interesting using the linkage based Dirichlet process mixture model to jointly estimate economic indicator curves through time from different countries and cluster countries based on the pattern of economic indicator curves.

Table 5.3: The estimated parameters by the linkage based Dirichlet process mixture model for 8 clusters are summarized.

Group	θ_1	θ_2	θ_3	θ_4	θ_5	σ^2
C_1	0.76	9.964	0.429	13.047	0.021	0.0018
	(0.744, 0.776)	(9.904, 10.023)	(0.417, 0.442)	(12.914, 13.253)	(0.02, 0.022)	(0.0015, 0.0023)
C_2	0.846	10.869	0.706	12.522	0.013	0.0006
	(0.834, 0.859)	(10.834, 10.903)	(0.688, 0.724)	(11.817, 13.141)	(0.012, 0.013)	(0.0005, 0.0008)
C_3	0.879	13.898	0.597	15.822	0.011	0.0014
	(0.861, 0.898)	(13.853, 13.945)	(0.571, 0.623)	(15.111, 16.387)	(0.011, 0.012)	(0.001, 0.0018)
C_4	0.886	12.327	0.787	14.96	0.009	0.0007
	(0.875, 0.896)	(12.295, 12.358)	(0.763, 0.811)	(14.391, 15.458)	(0.008, 0.009)	(0.0005, 0.001)
C_5	0.564	13.967	0.785	15.259	0.031	0.0017
	(0.52, 0.6)	(13.828, 14.092)	(0.687, 0.898)	(14.569, 16.252)	(0.03, 0.032)	(0.0009, 0.0032)
C_6	0.689	6.296	0.714	9.181	0.039	0.0014
	(0.641, 0.74)	(6.167, 6.427)	(0.663, 0.776)	(8.629, 9.844)	(0.033, 0.044)	(0.0007, 0.0027)
C_7	0.445	9.689	0.142	14.427	0.049	0.0036
	(0.372, 0.533)	(8.327, 11.286)	(0.119, 0.169)	(14.006, 14.887)	(0.047, 0.05)	(0.0019, 0.0069)
C_8	0.823	8.494	0.698	11.133	0.017	0.001
	(0.8, 0.846)	(8.418, 8.567)	(0.665, 0.733)	(10.444, 11.918)	(0.015, 0.019)	(0.0006, 0.0015)

Table 5.4: Cluster assignments for 30 building construction projects by the linkage based Dirichlet process mixture model.

Cluster	Building projects
C_1	Basketball Practice Facility, Biology Building/Vivarium Facility, Boiler Pollution Control Improvement, Main Campus Chilled Water Plant Add, ,Classroom Improvements - Phase I, Graduate School Facility, Substation Expansion, Upper Quad Conversion.
C_2	Chemistry/Physics - Phase II, Cowgill Hall HVAC and Power, Dietrick Servery/HVAC - Phase II, Hampton AREC Wing Replacement, Inst of Critical Tech and Applied Sci, Litton Reaves Hall Exterior Repairs, Surge Space Building.
C_3	Agriculture/Natural Resources Lab,Addition To Cheatham Hall, Fisheries and Aquatics Research Ctr, Add'l Recreat'n/Counseling/Clinical, Electric Service Facility.
C_4	Alumni/CEC/Hotel Complex, Bishop-Favrao/Bldg Construction Lab, Dairy Science Facilities, Geotechnical Lab, Multi-Purpose Livestock Arena.
C_5	Infectious Waste Incinerator 15232.
C_6	Football Fields.
C_7	Recreation Fields.
C_8	Career Services Facility, New Residence Hall.



(a) θ_1 (b) θ_2

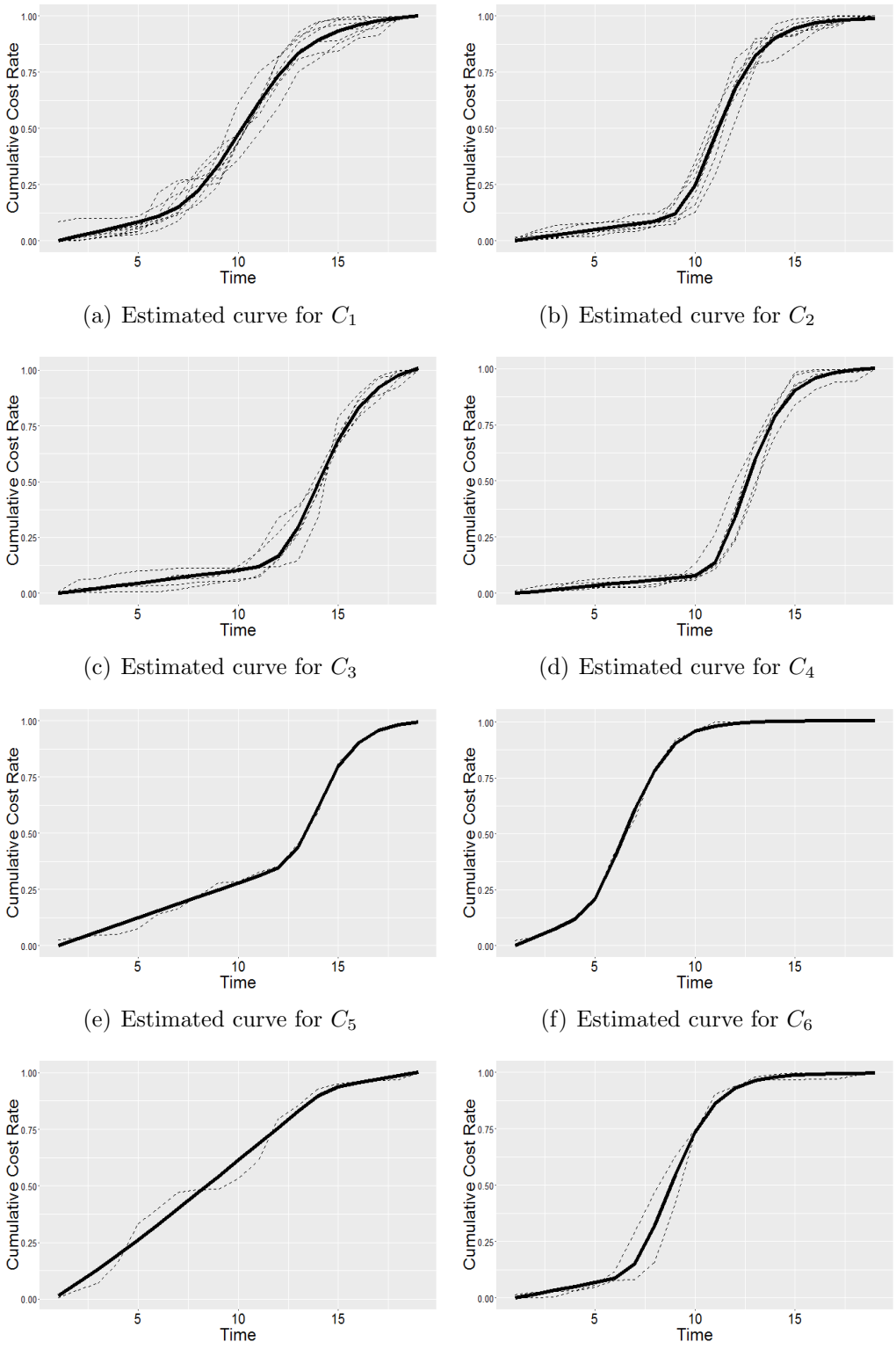
(c) θ_3 (d) θ_4

(e) θ_5 (f) σ^2

Figure 5.9: Panels (a), (b), (c), (d), (e), and (f) depict the distributions of the MCMC samples for θ_1, θ_2, θ_3, θ_4, θ_5, and σ^2 in Equation 5.2,respectively. C_1, C_2,. . ., and C_8 represent the eight clusters defined by LB-DP and D_1, D_2,. . ., and D_9 represent the nine clusters defined by DP-MBJ.

Table 5.5: Cluster assignments for 30 building construction projects by DP-MBJ.

Cluster	Building projects
D_1	Basketball Practice Facility, Biology Building/Vivarium Facility, Boiler Pollution Control Improvement, Main Campus Chilled Water Plant Add, ,Classroom Improvements - Phase I, Graduate School Facility, Substation Expansion, Upper Quad Conversion.
D_2	Chemistry/Physics - Phase II, Cowgill Hall HVAC and Power, Dietrick Servery/HVAC - Phase II, Hampton AREC Wing Replacement, Inst of Critical Tech and Applied Sci, Litton Reaves Hall Exterior Repairs, Surge Space Building.
D_3	Agriculture/Natural Resources Lab,Addition To Cheatham Hall, Fisheries and Aquatics Research Ctr, Add'l Recreat'n/Counseling/Clinical, Electric Service Facility.
D_4	Alumni/CEC/Hotel Complex, Bishop-Favrao/Bldg Construction Lab, Dairy Science Facilities, Geotechnical Lab, Multi-Purpose Livestock Arena.
D_5	Infectious Waste Incinerator 15232.
D_6	Football Fields.
D_7	Recreation Fields.
D_8	Career Services Facility.
D_9	New Residence Hall.



(a) Estimated curve for C_1

(b) Estimated curve for C_2

(c) Estimated curve for C_3

(d) Estimated curve for C_4

(e) Estimated curve for C_5

(f) Estimated curve for C_6

(g) Estimated curve for C_7

(h) Estimated curve for C_8

Figure 5.10: 8 estimated curves by LB-DP:HD are illustrated with their members.

5.2 Conclusion

As an application of linkage based Dirichlet process mixture model, we have modeled cumulative expenditure rate curves of building construction projects at Virginia Tech. Our findings are that the linkage based Dirichlet process has the capacity to find a reasonable cluster structure given observed data, and estimate the concentration parameter corresponding to the appropriate number of clusters. Compared to DP-MBJ, the linkage based Dirichlet process mixture model has provided the clustering solutions with a better silhouette coefficient for modeling the timeline for building construction costs.

Chapter 6

Linkage Based Nested Dirichlet Processes

As previously proposed, linkage based Dirichlet processes yield an approach for estimating the concentration parameter α in DPs. Extensions of Dirichlet processes include but are not limited to the following: hierarchical Dirichlet processes (Teh et al., 2006), dependent Dirichlet processes (MacEachern, 2000), and nested Dirichlet processes (Rodriguez et al., 2008). In this work, we particularly extend linkage based Dirichlet processes to the nested Dirichlet process setting. First, we introduce our motivation and review nested Dirichlet processes briefly. Then, we propose our extension, which is named linkage based nested Dirichlet processes.

6.1 Motivation

We have stated that Dirichlet process mixture models have the capability of accommodating an infinite number of mixture components and have the strength wherein users do not need

to specify the prior information on the number of components in advance. Most applications of Dirichlet processes rest on the underlying assumption that exchangeable samples are from an unknown distribution (MacEachern and Müller, 1998; Neal, 2000). There has been an increased need for extending Dirichlet processes when the structure of data is not simple but is nested or hierarchical, because we cannot posit the underlying exchangeability assumption (Teh et al., 2006; Rodriguez et al., 2008). Consider that data is nested in groups. For example, an academic university has different divisions such as a the college of science, a college of engineering, etc. Then, each division comprises various departments, for instance, the department of statistics and the department of mathematics in the college of science. Suppose that we would like to cluster based on students' GPA from the University. Observed GPAs from the University are nested in two layers: division level and department level. In this case, rather than using a classical Dirichlet process prior for clustering, other types of Dirichlet processes that allow having complex data structures seem more appropriate. Nested Dirichlet processes, hierarchical Dirichlet processes, and dependent Dirichlet processes are options for dealing with complex data (MacEachern, 2000; Teh et al., 2006; Rodriguez et al., 2008).

In this work, we propose linkage based nested Dirichlet process, which is an extension of a linkage based Dirichlet process and an alternative to the nested Dirichlet process. This work is motivated by two concentration parameters (α and β) in nested Dirichlet processes that have an effect on estimating the number of global clusters (or centers) and the number of local clusters (or sub-clusters). So far, there have been few studies on choosing the concentration parameters in the nested Dirichlet processes (Rodriguez et al., 2008). Rodriguez et al. (2008) suggests employing Gamma priors on both concentration parameters or having the values of the concentration parameters fixed. Also, given K segmented centers, Rodriguez et al. (2008) has only one concentration parameter (β) for defining sub-clusters within K segmented centers. Unlike nested Dirichlet processes, our linkage based nested Dirichlet process allows us to have an equal number of the concentration parameters (β_1,

$\beta_2,\ldots,\ \beta_K$) to the number of the segmented centers. The rest of this section is organized as follows: we first review the nested Dirichlet process, introduce how to extend our proposed method to nested Dirichlet processes, and show how to implement a linkage based nested Dirichlet process in the Markov Chain Monte Carlo algorithm.

6.2 Nested Dirichlet Processes

The nested Dirichlet process introduced by Rodriguez et al. (2008) broadens the scope of using Dirichlet processes. Particularly, mixture models with the nested Dirichlet process are beneficial for multi-level clustering analyses because this stochastic process enables us to cluster probability distributions, which are the mixture of sub-clusters, and to configure the sub-clusters. For example, consider SAT scores of students from different universities. The probability distribution of SAT scores within a particular university can be assumed to be the mixture of sub-clusters because most universities value student diversity in their universities. Also, because universities have different qualifications for entrance, the probability distributions of universities based on their SAT scores of students might be distinct from each other. Thus, clustering universities-according to their probability distribution of SAT-should be considered for identifying unique universities. Clustering universities and grouping students within grouped universities can be accomplished by the use of the nested Dirichlet process.

Rodriguez et al. (2008) defines the nested Dirichlet process as follows. For $j = 1, 2, \ldots, J$ and $i = 1, 2, \ldots, n_j$, we assume that observations $\mathbf{y_j} = (y_{1j}, y_{2j}, \ldots, y_{n_j j})$ within distribution j are exchangeable and $\mathbf{y_j} \sim F_j$. As an example, we can consider y_{ij} the SAT score of student i in university j. Given a collection of mixing distribution, $\{G_1, G_2, \ldots, G_J\}$, let K be the unique number of clustered centers and G_k^* be the distribution for grouped centers into k for

$k = 1, 2, \ldots, K$. Then, the random distributions $F_1, F_2, \ldots, F_J$ are as follows:

$$F_j(\mid \phi) = \int_{\Theta} p(. \mid \theta, \phi) G_j(d\theta),$$
$$G_j \equiv \sum_{k=1}^{\infty} \pi_k^* \delta_{G_k^*},$$
$$G_k^* \equiv \sum_{l=1}^{\infty} \omega_{lk}^* \delta_{\theta_{lk}^*},$$

where θ and ϕ are the parameters in the distribution p, π_k is the mixing proportion for k distribution and ω_{lk} is the mixing proportion for l sub-cluster in k distribution, such that $\sum \pi_k = 1$ and $\sum \omega_{lk} = 1$. Then, the mixture model with the nested Dirichlet process can be formed as follows:

$$y_{ij} \sim p(y_{ij} \mid \theta_{ij}, \phi),$$
$$\theta_{ij} \sim G_j,$$
$$G_1, G_2, \ldots, G_J \sim DP(\alpha DP(\beta, G_0)),$$

where α and β are the concentration parameters and G_0 is the base measure in the nested Dirichlet process. In the nested Dirichlet process, there are two concentration parameters, α and β, which determine the number of unique distributions (or centers) and the number of sub-clusters within unique distributions, respectively. Similar to $P(\theta_j = \theta_{j'}) = \frac{1}{1+\alpha}$ in the DP, the nested Dirichlet ensures that $P(G_j = G_{j'}) = \frac{1}{1+\alpha}$ for the mixing distribution $G = \{G_1, G_2, \ldots, G_J\}$. Also, by the stick breaking construction, the marginal distribution of each G_j follows the Dirichlet process with β and H (Rodriguez et al., 2008).

To model the nested Dirichlet process, we start with a stick-breaking representation of the Dirichlet process. Intuitively, if we replace the random sample from the Dirichlet process with the random probability measure, a nested Dirichlet process prior results. In other words, the nested Dirichlet process is the collection of distributions for different centers.

The stick breaking process construction for nested Dirichlet processes is in shown Figure 6.1.

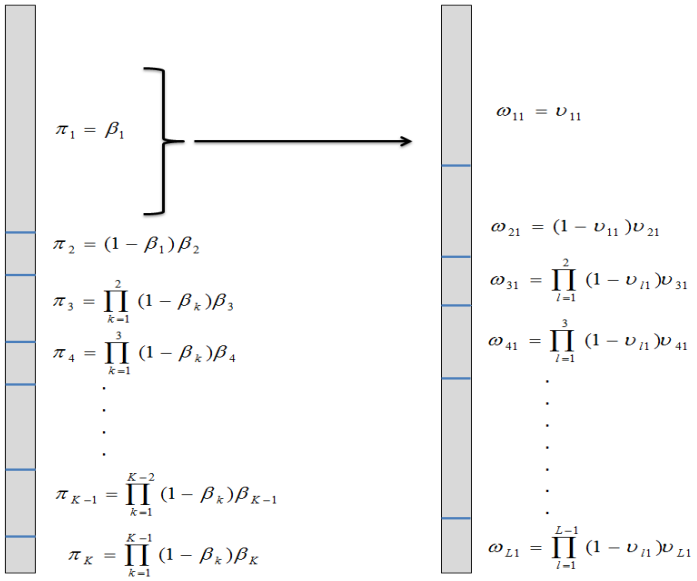
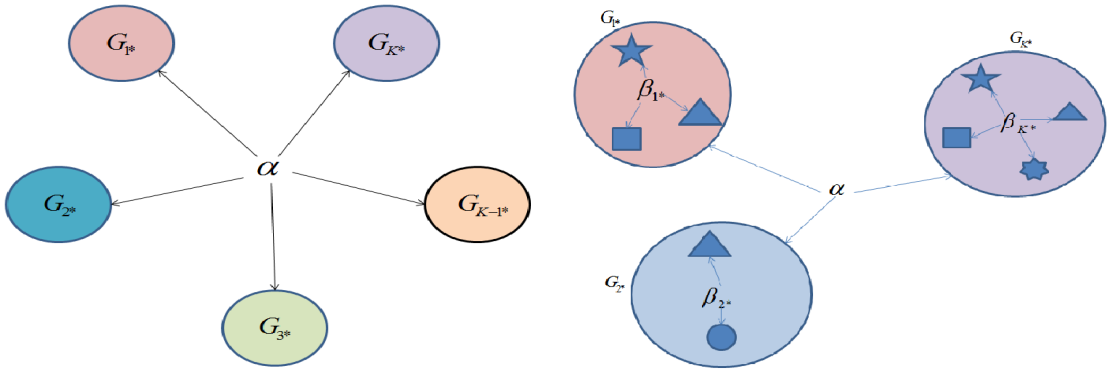


Figure 6.1: An illustration of a stick-breaking process for a nested Dirichlet process. When $K \to \infty$ and $L \to \infty$, the stick-breaking process becomes the nested Dirichlet process.

6.3 Linkage Based Nested Dirichlet Processes

A linkage based nested Dirichlet process is motivated by the estimation of the concentration parameter β for defining sub-clusters within distributions. For typical nested Dirichlet processes, one β has been used for defining sub-clusters within distributions. However, we expect the shapes of sub-clusters and number of sub-clusters in $\{G_1^*, G_2^*, \ldots, G_K^*\}$ are different from each other. Thus, we suggest using different βs for K grouped centers in order to enhance the performance of clustering, especially for defining meaningful sub-groups. In other words, our motivation for linkage based nested Dirichlet processes starts from the idea that we should use $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_K)$ for defining sub-clusters within grouped centers.

For convenience, denote the distribution G_k^* and the sub-cluster l within G_k^* by C_k and



(a) α for distributions (b) $\boldsymbol{\beta}$ for sub-clusters within distributions

Figure 6.2: Our motivation of the linkage based nested Dirichlet process. Panels (a) and (b) describe the motivation of estimating the concentration parameter for the number of distributions and the concentration parameters for the number of sub-clusters in clustered distributions, respectively.

C_{lk}. To construct the linkage based nested Dirichlet process, the function in Equation 3.8 is revisited for C_k and C_{lk} for $k = 1, 2, \ldots, K$ and $l = 1, 2, \ldots, L$:

$$\begin{aligned} S(C, K) &= \binom{K}{2} - \sum_{i<j}^{K} D(C_i, C_j), \\ S(C_k, L) &= \binom{L}{2} - \sum_{i<j}^{L} D(C_{ik}, C_{jk}). \end{aligned} \tag{6.1}$$

We estimate $K + 1$ concentration parameters, α and $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_K\}$, by measuring distances between clustered distributions and distances between sub-clusters within each clustered distributions. The solution for the optimal number of distributions is the same as the solution in Equation 3.10; the estimate for α is the solution by solving Equation 3.11 with respect to α, same as the one in the linkage based Dirichlet process. For $k = 1, \ldots, K$, the expected number of sub-clusters for k distribution is as follows:

$$E(L \mid \beta_k, n_k) = \sum_{i=1}^{n_k} \frac{\beta_k}{\beta_k + i - 1}. \tag{6.2}$$

To update $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$, we define the optimal number of sub-clusters L'_k for k clustered distributions as follows:

$$L'_k = \frac{1 + (1 + 8 \sum_{i<j}^{L_k} D(C_{ik}, C_{jk}))^{1/2}}{2}. \tag{6.3}$$

We replace $E(L_k \,|\, \beta_k,\, n_k)$ with L'_k in Equation 6.3. We rewrite Equation 6.2 as follows:

$$L'_k = \sum_{i=1}^{n_k} \frac{\beta_k}{\beta_k + i - 1} \tag{6.4}$$

Then, we solve Equation 6.4 with respect to β_k. For the linkage based nested Dirichlet process, we can write our model as follows:

$$\begin{aligned} y_{ij} &\sim p(y_{ij} \mid \theta_{ij}), \\ \theta_{ij} &\sim G_j, \\ G_j &\sim \text{LBNDP}(\alpha, \boldsymbol{\beta}, H), \end{aligned} \tag{6.5}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_J)$. In addition, we calculate a covariance between random distributions $\{G_1, G_2, \ldots, G_J\}$ in the linkage based nested Dirichlet process as follows:

$$Cov(G_i(B), G_j(B)) = \begin{cases} \frac{H(B)(1-H(B))}{\beta+1} & \text{if} \quad G_i = G_j \\ 0 & \text{if} \quad G_i \neq G_j \end{cases}, \tag{6.6}$$

for any finite and measurable partition B of a measurable space Ω. In Equation 6.6, we can induce $\beta = \beta_i = \beta_j$ if $G_i = G_j$, then the covariance between two random distribution G_i and G_j becomes $Var(G_j(B))$ for each j. However, if $G_i \neq G_j$, the covariance between G_i and G_j becomes 0 in the linkage based nested Dirichlet process. Equation 6.6 expresses no covariance between observations from different distributions but a positive covariance between the observations in the same distribution. This is in line with our initial idea of

both linkage Dirichlet process prior and linkage based nested Dirichlet process prior, because our goal of clustering analyses is identifying distinct distributions.

6.3.1 Extension of DP-MBJ to Nested Dirichlet Processes

We extend DP-MBJ to the nested Dirichlet process setting so that we can compare our linkage based nested Dirichlet process. The logic for extending DP-MBJ is simple. We estimate the same number of concentration parameters as the linkage based nested Dirichlet process estimates; the step to estimate the concentration parameter α for grouped distributions is same as in DP-MBJ. The way to estimate the concentration parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_K)$ is as follows:

$$\frac{1}{B}\sum_{b=1}^{B} L_k = \sum_{i=1}^{n_k} \frac{\beta_k}{\beta_k + i - 1}, \tag{6.7}$$

where L_k is the number of sub-clusters in k distribution. We repeat this process for $k = \{1, 2, \ldots, K\}$. We can obtain B samples of L_k from the MCMC chain and solve Equation 6.7 with respect to β_k.

6.3.2 Gibbs Sampling Implementation

The MCMC sampling scheme for nested Dirichlet processes is introduced in Rodriguez et al. (2008). The scheme in Rodriguez et al. (2008) uses the idea from Ishwaran and James (2001), which is the sampler that replaces the infinite sum by a finite sum. The sampling scheme for linkage based nested Dirichlet processes is introduced in Algorithm 2.

6.3.3 Property of Linkage Based Nested Dirichlet Processes

Let $\{G_1(B), G_2(B), G_3(B), \cdots, G_J(B)\}$ be the collection of random variables, then, the correlation between $G_i(B)$ and $G_j(B)$ for Linkage Based Nested Dirichlet processes is as

Algorithm 2 MCMC algorithm for LB-NDP mixture models.

Initialize all parameters and choose K and L for double truncation.
for $t = 1$ to T **do**
  for $k = 1$ to K **do**
    Sample the center indicators ζ_j for $j = 1, 2, \ldots, J$ with

$$p(\zeta_j = k \mid .) = \pi_k^{(t-1)} \prod_{i=1}^{n_j} \sum_{l=1}^{L} \omega_{lk}^{(t-1)} p(y_{ij} \mid \theta_{lk}^{(t-1)}, \phi^{(t-1)}).$$

    Sample $u_k \sim beta(1 + m_k, \alpha^{(t-1)} + \sum_{s=k+1}^{K} m_s)$ given the number of centers in k, m_k.
    Update $\pi_k = u_k \prod_{s=1}^{k-1}(1 - u_s)$.
    Construct a distance matrix D for unique $K*$ centers.
    Update $\hat{K}$ such that

$$\hat{K} = \frac{1 + \sqrt{1 + 8 \sum_{i<j}^{K^*} D(C_i, C_j)}}{2}.$$

    Estimate $\alpha^{(t)}$ which satisfies $\hat{K} = \sum_{j=1}^{J} \frac{\alpha}{\alpha + j - 1}$.
    for $l = 1$ to L **do**
      Draw the sub-cluster indicator ξ_{ij} for $i = 1, 2, \ldots, n_j$ with

$$p(\xi_{ik} = l \mid .) \propto w_{l\zeta_k}^{(t-1)} p(y_{ik} \mid \theta_{l\zeta_k}^{(t-1)}, \phi^{(t-1)}).$$

      Sample $\nu_{lk} \sim beta(1 + n_{lk}, \beta_k^{(t-1)} + \sum_{s=l+1}^{L} n_{ls})$, where n_{lk} is the number of obs in l sub-cluster in k center.
      Update $\omega_{lk} = \nu_{lk} \prod_{s=1}^{l-1}(1 - \nu_{sk})$.
      Construct the distance matrix D_{k*} for sub-clusters in $k*$ centers.
      Update $\hat{L_{k*}}$ such that

$$\hat{L_{k*}} = \frac{1 + \sqrt{1 + 8 \sum_{i<j}^{l_{k*}} D_k(C_{ik*}, C_{jk*})}}{2}.$$

      Estimate $\beta_k^{(t)}$ which satisfies $\hat{L_{k*}} = \sum_{i=1}^{n_{k*}} \frac{\beta_k}{\beta_k + i - 1}$.
      Draw a new sample for $\theta_{l\zeta_k}^{(t)} \sim p(\theta_{l\zeta_k}^{(t-1)} \mid y_{ik})$.
    end for l
  end for k
  Draw a new sample for $\phi^{(t)} \sim \prod_{J}^{j=1} \prod_{n_j}^{i=1} p(y_{ij} \mid \theta_{l\zeta_k}^{(t)}, \phi) p(\phi)$.
end for t

follows: If G_i and G_j are equal, then let $\beta = \beta_i = \beta_j$.

$$\begin{aligned} Cov(G_i(B), G_j(B) \mid G_i = G_j) &= E(G_i(B)G_j(B) \mid G_i = G_j) \\ &\quad - E(G_i(B) \mid G_i = G_j)E(G_j(B) \mid G_i = G_j) \\ &= E(G_i^2(B)) - E^2(G_i(B)) \\ &= Var(G_i(B)) \\ &= \frac{H(B)(1 - H(B))}{\beta + 1} \end{aligned} \tag{6.8}$$

For $G_i \neq G_j$,

$$\begin{aligned} Cov(G_i(B), G_j(B) \mid G_i \neq G_j) &= E(G_i(B)G_j(B) \mid G_i \neq G_j) \\ &\quad - E(G_i(B) \mid G_i \neq G_j)E(G_j(B) \mid G_i \neq G_j) \\ &= E(G_i(B))E(G_j(B)) - E(G_i(B))E(G_j(B)) \\ &= 0 \end{aligned} \tag{6.9}$$

Thus,

$$\begin{aligned} Cov(G_i(B), G_j(B)) &= Cov(G_i(B), G_j(B) \mid G_i = G_j)p(G_i = G_j) \\ &\quad + Cov(G_i(B), G_j(B) \mid G_i \neq G_j)p(G_i \neq G_j) \\ &= \frac{H(B)(1 - H(B))}{\beta + 1}\frac{1}{\alpha + 1} + 0\frac{\alpha}{\alpha + 1}. \end{aligned} \tag{6.10}$$

Chapter 7

Linkage Based Nested Dirichlet Processes: Simulation and Application

In Chapter 7, we illustrate a set of simulation studies for investigating the performance of the linkage based nested Dirichlet process, which is the extension of the linkage based Dirichlet process when the data is nested in groups. In addition, by modeling the median household income data in the United States, we cluster the states first and then segment the counties within the clustered states into the optimal number of groups. We will compare the modeling result obtained by the linkage based nested Dirichlet process to the clustering result defined by the extension of DP-MBJ, which we refer to as NDP-MBJ.

7.1 Simulation

In this section, we perform a sequence of simulation studies to verify the performance of linkage based nested Dirichlet processes. Like the linkage based nested Dirichlet process, we

also extend DP-MBJ to the nested Dirichlet process as shown in Chapter 6. Specifically, DP-MBJ estimates the concentration parameters (β_1, β_2, ..., β_K) by obtaining B samples of the number of sub-clusters belonging to each K segmented centers. In this section, we introduce our simulation designs and the scheme to quantify the performance of the linkage based nested Dirichlet processes as an application of the multi-level clustering algorithm. Then, we demonstrate our simulation results by comparing with the performance of NDP-MBJ.

7.1.1 Simulation Design

We present our simulation design to generate the distributions that comprising a mixture of Gaussian components. The simulation set up is as follows: we generate K distributions of each size n from the mixtures of J Gaussian components. For each distribution with J Gaussian components, we start by sampling the mixing proportion $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_J)$ from a Dirichlet distribution. For each distribution, the number of Gaussian components J can differ. In detail, we show how we generate the distribution of the mixture of univariate Gaussians:

Step 1. Choose the number of mixture components J arbitrarily.

Step 2. Given J, sample $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_J)$ from $Dir(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_J)$.

Step 3. Choose $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_J)$.

Step 4. Generate $\boldsymbol{y} = \{y_1, y_2, \ldots, y_n\}$ from $p(\boldsymbol{y} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \sigma^2) = \sum_j \pi_j p(\boldsymbol{y} \mid \mu_j, \sigma^2)$.

In simulation studies, we choose $K = 6$ and generate six distributions, i.e. D_1, D_2, D_3, D_4, D_5, and D_6. We intentionally plot a scenario where D_1 and D_4, D_2 and D_5, and D_3 and D_6 share the same Gaussian components. However, the mixture weights are different in order to see not only whether LB-NDP has the capability of clustering the distribution of the mixtures but also to measure the effect of estimated concentration parameters when the size of n differs. Across the simulation study, we use $n = \{20, 50, 100\}$ in order to assess

the the influence of the sample size on the concentration parameter estimation and modeling results. We estimate the concentration parameter using the linkage based nested Dirichlet process and the extension of DP-MBJ (NDP-MBJ). For sampling the mixture weights from the Dirichlet distribution, we choose $\alpha = 1$. When generating data points, we fix σ at 0.5 for the distribution of the mixture of the univariate Gaussians. For bivariate Gaussians, we use $\Sigma = 0.5^2 I$. We use the double truncation approximation (Rodriguez et al. (2008)) for easier computation. We choose the truncation level $K = 10$ for distributions and the truncation level $J = 10$ for sub-clusters.

There might be several approaches to summarize the simulation results. For example, calculating KL divergence of the density estimates to the true densities may be the one way to summarize the results. However, we are more interested in measuring the performance of LB-NDP as an application of multi-level clustering analysis. Thus, we will compare the modeling results using the overall silhouette coefficient. Since each segmented cluster via LB-NDP/NDP-MBJ is consisted of the sub-clusters, we obtain the silhouette coefficients from sub-clusters within each segmented distributions. We then call the average of these silhouette coefficients the overall silhouette coefficient to quantify the overall clustering qualities. Also, when we calculate the distances between distributions and the distances between sub-clusters, we utilize the total variation distance:

$$TVD(f, g) \approx \frac{1}{2} \sum_{i=1}^{N} |f(x_i) - g(x_i)|, \tag{7.1}$$

where $N = \sum_{k=1}^{K} n_k$.

Non-overlapping distributions from mixtures of univariate Gaussians distributions

In our simulation study, we simulate six distributions from the mixture of univariate Gaussians. We allow D_1 and D_4, D_2 and D_5, and D_3 and D_6 to have the same Gaussian components, but different weights. However, when we generate Gaussian components for these distributions, we purposely separate these three pairs (D_1,D_4), (D_2,D_5), and (D_3,D_6) from each other, so that these pairs are unlikely to share the area under density curves as shown in Figure 7.1. As a base measure for both LB-NDP and NDP-MBJ, we choose a normal inverse gamma distribution, $NIG(0, 0.01, 1, 2)$. This simulation aims to investigate whether LB-NDP keeps separating these three pairs from each other and provides the better clustering solution by defining the optimal number of sub-clusters than NDP-MBJ based on the calculation of the overall silhouette coefficient. For each n, we repeat the experiments 200 times. All results are based on $5,000$ MCMC samples obtained after $20,000$ iterations. We use the double truncation approximation (Rodriguez et al. (2008)). We choose the truncation level $K = 10$ for distributions and the truncation level $J = 10$ for sub-clusters.

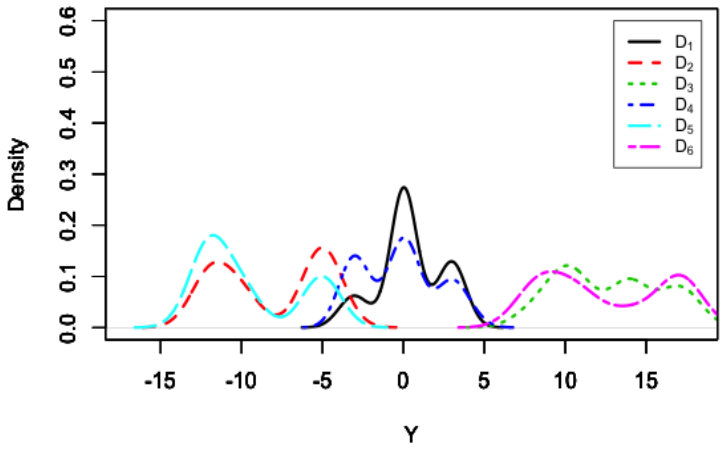


Figure 7.1: An example of generated distributions for non-overlapping distributions from mixtures of univariate Gaussians case.

Overlapping distributions from mixtures of univariate Gaussians case

Unlike non-overlapping distributions from the mixtures of univariate Gaussians case, (D_1,D_4) partially shares the Gaussian components with both (D_2,D_5) and (D_3,D_6), as shown in Figure 7.2. We use a normal inverse gamma distribution, $NIG(0, 0.01, 1, 2)$ as our base measure in LB-NDP and NDP-MBJ. This simulation study investigates whether LB-NDP reinforces D_1 and D_4 to be merged into other distributions and finds the better clustering solution than NDP-MBJ by obtaining the overall silhouette coefficients. Also, for each n, we repeat the experiments 200 times. The simulation results are summarized based on based on $5,000$ MCMC samples obtained after $20,000$ iterations.

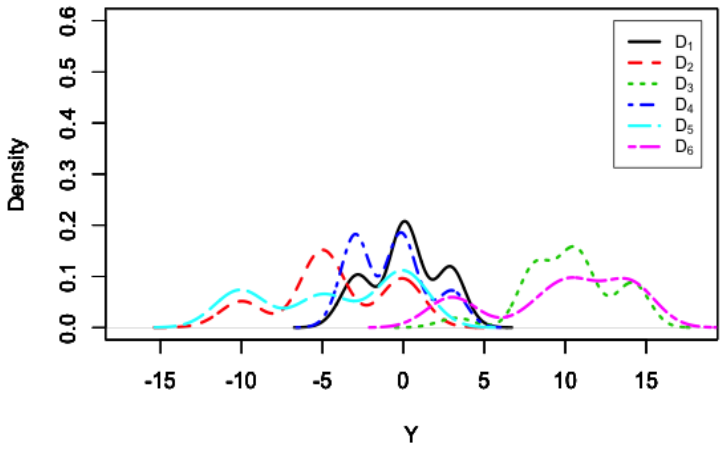


Figure 7.2: An example of simulated distributions for overlapping distributions from mixtures of univariate Gaussians case.

Non-overlapping distributions from mixtures of bivariate Gaussians case

We extend our simulation study to model non-overlapping distributions from mixtures of bivariate Gaussians distributions. The simulation set up is same as the design for univariate

distributions. We designate D_1 and D_4, D_2 and D_5, and D_3 and D_6 to share the same bivariate Gaussian components but weighted differently. For each n, we generate 200 different data sets and summarize the simulation results based on $10,000$ samples after $20,000$ iterations in the MCMC chain.

Overlapping distributions from mixtures of bivariate Gaussians case

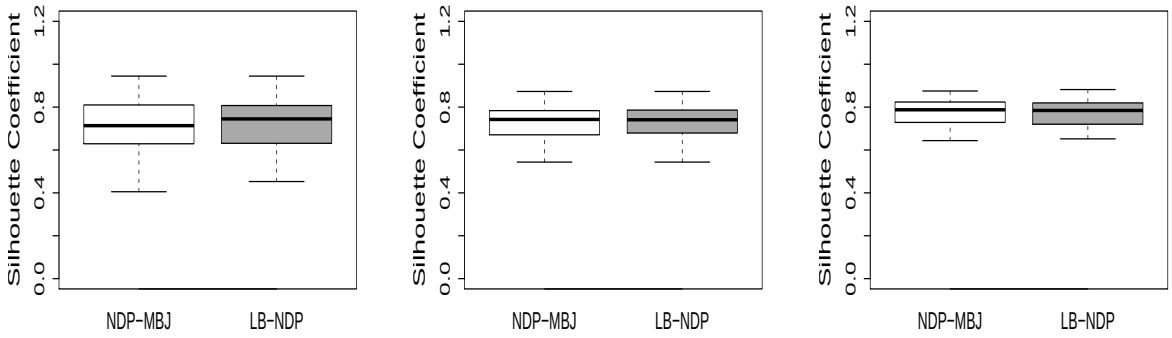
We simulate six distributions from mixtures of bivariate Gaussians. The set up for the simulation as follows: we design (D_1,D_4) to share the piece of the Gaussian components with both (D_2,D_5) and (D_3,D_6). This simulation study aims to analyze the performance of LB-NDP over NDP-MBJ when there are distributions overlapping each other. We obtain the overall silhouette coefficients based on $10,000$ samples after the burn-in period to summarize the simulation results for each n.

7.1.2 Simulation Results

We report the experiment's results based on the different sets of data with different size of $n = \{20, 50, 100\}$, which is the number of data points for each distribution. Also, recall that our number of simulated distributions is equal to six. This means for $n = 20$, we generate 120 data points in total. We choose the numbers of mixture components for six distributions are 3, 5, 3, 3, 5, and 3, respectively. Since we now measure the distance between the distribution of the mixtures, we are not able to use the Hellinger distance in Equation 3.4 for bivariate Gaussian cases. Thus, we calculate the total variation distance by utilizing kernel density estimation to measure the proximity between distributions and the proximity between sub-clusters. We use $B = 100$ for implementing NDP-MBJ.

Non-overlapping distributions from mixtures of univariate Gaussians case

We demonstrate the simulation result for non-overlapping distributions from mixtures of univariate Gaussians case, where the number of data points for each distribution is $n = \{20, 50, 100\}$. Figure 7.3 displays the distributions of the overall silhouette coefficients obtained by applying NDP-MBJ and LB-NDP. We can verify that the distribution of the overall silhouette coefficients for NDP-MBJ (white) and the distribution of the overall silhouette coefficients for LB-NDP (dark-gray) look similar to each other when $n = 50$ and $n = 100$. However, as depicted in panel (a) in Figure 7.3, LB-NDP shows the higher median of the overall silhouette coefficients than NDP-MBJ when $n = 20$. In theory, the effect of the concentration parameter is considerable given the small number of observations. We have examined this via the simulation studies in Chapter 4. As the overall silhouette coefficient is based on the silhouette coefficients from sub-clusters within segmented distributions, we can see that the estimation of concentration parameters, $\beta_1, \beta_2, \ldots, \beta_K$, for K distributions influences the clustering results when the number of observations within the distribution is small.

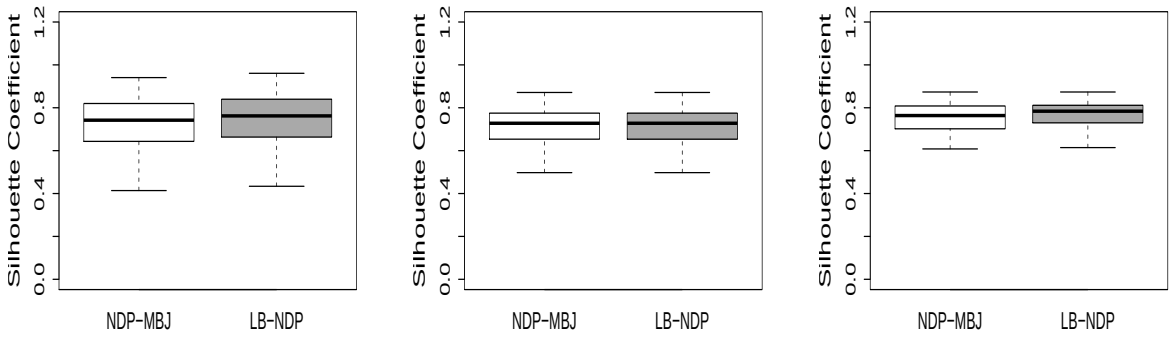


(a) n=20 (b) n=50 (c) n=100

Figure 7.3: Panels (a), (b), and (c) illustrate the distributions of the overall silhouette coefficients for $n = 20$, $n = 50$, and $n = 100$ for non-overlapping distributions respectively.

Overlapping distributions from mixtures of univariate Gaussians case

We illustrate the simulation result for overlapping distributions from mixtures of univariate Gaussians case given $n = \{20, 50, 100\}$. Figure 7.4 depicts the distributions of the overall silhouette coefficients obtained by applying NDP-MBJ (white) and LB-NDP (dark-gray). As shown in panels (a), (b), and (c) in Figure 7.4, across the size of n, we can see that the median of the overall silhouette coefficients obtained by LB-NDP is consistently and slightly higher than the median of the overall silhouette coefficients obtained by NDP-MBJ. Also, we can see that there is less variability in the overall silhouette coefficients for LB-NDP than NDP-MBJ. We conclude that when there are distributions overlapping with other distributions across the range of data points, LB-NDP has better performance than NDP-MBJ because of the fact that LB-NDP merges the distributions/sub-clusters if they are similar to each other.

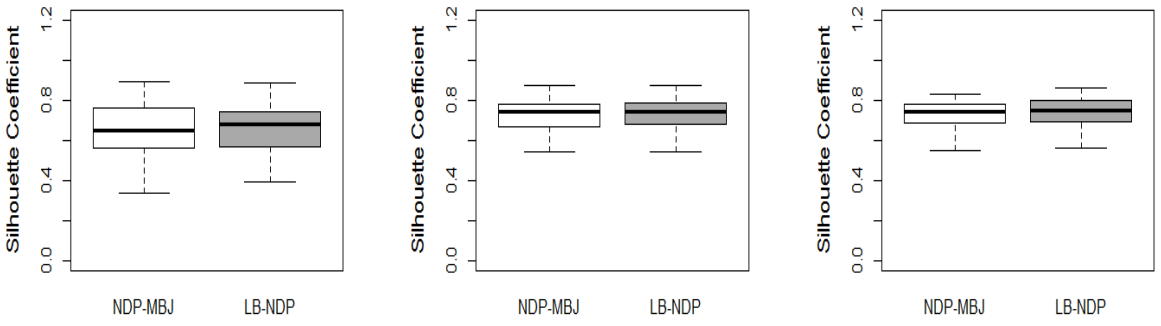


(a) n=20 (b) n=50 (c) n=100

Figure 7.4: Panels (a), (b), and (c) illustrate the distributions of the overall silhouette coefficients for $n = 20$, $n = 50$, and $n = 100$ for overlapping distributions respectively.

Non-overlapping distributions from mixtures of bivariate Gaussians case

We provide the simulation result for non-overlapping distributions from mixtures of bivariate Gaussians case, where the number of data points for each distribution is $n = \{20, 50, 100\}$. Panel (a) in Figure 7.5 clearly shows LB-NDP has the higher median of the overall silhouette coefficients than NDP-MBJ when the number of data points for each distribution is 20. When $n = 50$ and $n = 100$, the distributions of the overall silhouette coefficients obtained by NDP-MBJ (white) and LB-NDP (dark-gray) look similar to each other. Like the simulation study for the univariate distributions, we can verify that the effect of the concentration parameter (α) is considerable when n is small. we can see that the estimation of concentration parameters, $\beta_1, \beta_2, \ldots, \beta_K$, for K distributions influences the clustering results when the number of observations within the distribution is small.

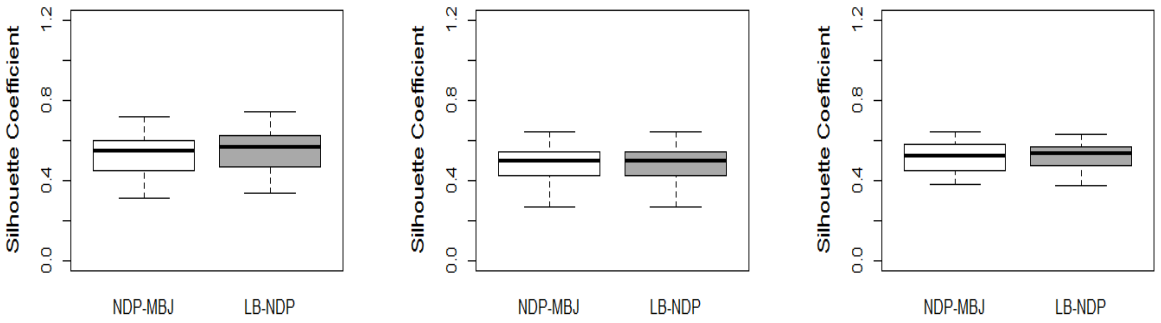


(a) n=20 (b) n=50 (c) n=100

Figure 7.5: Panels (a), (b), and (c) illustrate the distributions of the overall silhouette coefficients for $n = 20$, $n = 50$, and $n = 100$ for bivariate non-overlapping distributions respectively.

Overlapping distributions from mixtures of bivariate Gaussians case

We analyze the simulation result for overlapping distributions from mixtures of bivariate Gaussians when $n = \{20, 50, 100\}$. Figure 7.6 illustrates the distributions of the overall silhouette coefficients obtained by applying NDP-MBJ (white) and LB-NDP (dark-gray). As depicted in panel (a) in Figure 7.6, for bivariate Gaussians with $n = 50$, the median of the overall silhouette coefficients obtained by LB-NDP is slightly higher than the median of the overall silhouette coefficients obtained by NDP-MBJ. Also, we can see that both methods show similar levels of the performance. There is less variability in the overall silhouette coefficients for LB-NDP than NDP-MBJ. We conclude that when there are distributions overlapping with other distributions across the range of data points, LB-NDP has a better performance than NDP-MBJ because of the fact that LB-NDP merges the distributions/sub-clusters if they are similar to each other.



(a) n=20 (b) n=50 (c) n=100

Figure 7.6: Panels (a), (b), and (c) illustrate the distributions of the overall silhouette coefficients for $n = 20$, $n = 50$, and $n = 100$ for bivariate overlapping distributions respectively.

We have performed the simulation studies in order to analyze the performance of LB-NDP by comparing the performance of NDP-MBJ, which is the extension of DP-MBJ to the nested

Dirichlet process. As a metric to compare the performance, we use the overall silhouette coefficient, that is, the average of silhouette coefficients obtained from the segmented distributions. It is interesting that for overlapping distributions from the mixture of univariate Gaussian case, LB-NDP has performed better than NDP-MBJ across the size of n used in our simulation studies when the number of distributions is equal to six. We can infer that the reason is because LB-NDP, which uses the proximity information between distributions and between sub-clusters to estimate the concentration parameters, α, $\beta_1, \beta_2, \ldots, \beta_K$, forces a sub-cluster/distribution to be segmented into other sub-clusters/distributions, if they are close to each other.

7.2 Application: Modeling Median Household Income in the United States

In this section, we model the median household income in the United States. Using a linkage based nested Dirichlet process mixture model, we group the territories based on the distributions of the median household income from the counties in each state. The counties within the clustered states are then segmented into sub-clusters based on their median household income. We utilize the nested Dirichlet process mixture model with fixed concentration parameters and NDP-MBJ for comparing the clustering results.

7.2.1 Data

The median household income data is from the U.S. Census Bureau, 2009-2013 5-Year American Community Survey. The household income is a combined gross income of all people 15 years or older sharing the same house unit. Notably, the median household income is an economic statistic often used for comparing affluence and living standards between cities, counties, and states (DeNavas-Walt, 2010). Figure 7.7 depicts the states in the United States based on the median household income at the state level. Figure 7.7 does not show the distributions of the median household income collected from the counties in the states. In the original dataset, the collection of the median household income from counties and county equivalents in 51 territories states including the 50 states and the District of Columbia are collected. The number of counties per state differs; for example, 3 counties in Delaware, 254 counties in Texas, and 58 counties in California. We use 49 states that have more than 3 counties (excluding the District of Columbia and Delaware) and contain 3139 counties and county equivalents in total. Before analyzing the median household income data, we standardize the median household income.

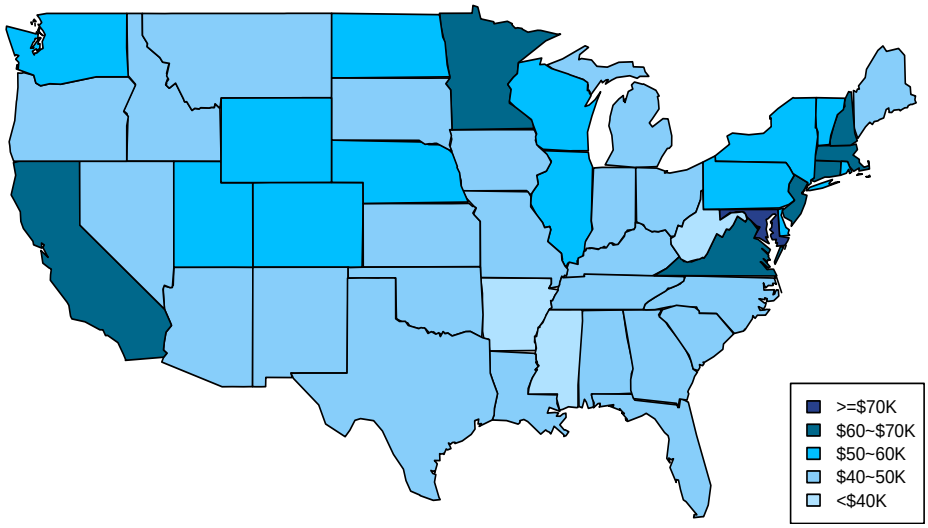


Figure 7.7: States in the United States are visualized according to their categorized median household income. Hawaii and Alaska are not depicted on the map.

7.2.2 Model Specification

The K-means clustering algorithm is not able to handle the data in the nested setting. Thus, Applying the K-means clustering algorithm to the median household data will bring the result that only groups counties and ignores the distribution of the median household income within the state. Therefore, we will not be able to see the differences in the distributions of the median household income by states. In order to model the median household data which are nested, we fit a nested Dirichlet process mixture model with fixed concentration parameters, a nested Dirichlet process mixture model with NDP-MBJ, and a linkage based nested Dirichlet process mixture model to the scaled median household income. Let y_{ij} be the scaled household income from i counties in state j. Then the nested Dirichlet process

mixture model, which we refer to as NDP, is as follows:

$$\begin{aligned} y_{ij} &\sim N(\theta_{ij}, \sigma_{ij}^2), \\ \theta_{ij} &\sim G_j, \\ G_1, G_2, \ldots, G_J &\sim DP(\alpha DP(\beta, G_0)), \end{aligned}$$

where $i \leq n_j$ and $j \leq J$. We use the fixed concentration parameters $\alpha = 1$ and $\beta = 1$. Also, we formulate the nested Dirichlet process mixture model with NDP-MBJ:

$$\begin{aligned} y_{ij} &\sim N(\theta_{ij}, \sigma_{ij}^2), \\ \theta_{ij} &\sim G_j, \\ G_1, G_2, \ldots, G_J &\sim DP(\alpha DP(\boldsymbol{\beta}, G_0)), \end{aligned}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_J)$. Our linkage based nested Dirichlet process mixture model for modeling the house median income is as follows:

$$\begin{aligned} y_{ij} &\sim N(\theta_{ij}, \sigma_{ij}^2), \\ \theta_{ij} &\sim G_j, \\ G_1, G_2, \ldots, G_J &\sim LB - NDP(\alpha, \boldsymbol{\beta}, G_0), \end{aligned}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_J)$. For efficient computation, we utilize the algorithm presented in Rodriguez et al. (2008), which employs the truncation approximation to the stick-breaking process. Let K be the truncation level at the level of states and L be the truncation level at the level of counties. We choose $K = 10$ and $L = 10$ for fitting three models. We use total variation distance to measure the distance between clustered states/counties in the linkage based nested Dirichlet process. We pick a normal inverse gamma distribution, $G_0 \sim NIG(0, 0.01, 1, 1)$ as our base function for three models.

As a preliminary study, we illustrate total variation distances between all pairs of states based on the distribution of the median household income in Figure 7.8. Figure 7.8 provides an insight into possible cluster structure. As shown in Figure 7.8, we define a distinct cluster structure of the states on the lower left corner of Figure 7.8. For example, we can see that MD, NH, AK, MA, CT, HI, NJ, RI are well-separated from other states. However, it looks like a heterogeneous partition of the states also exists, such as states CA and NV. This may provide the insight to obtain the various number of distinct partitions of the states depending on the choice of the concentration parameter α.

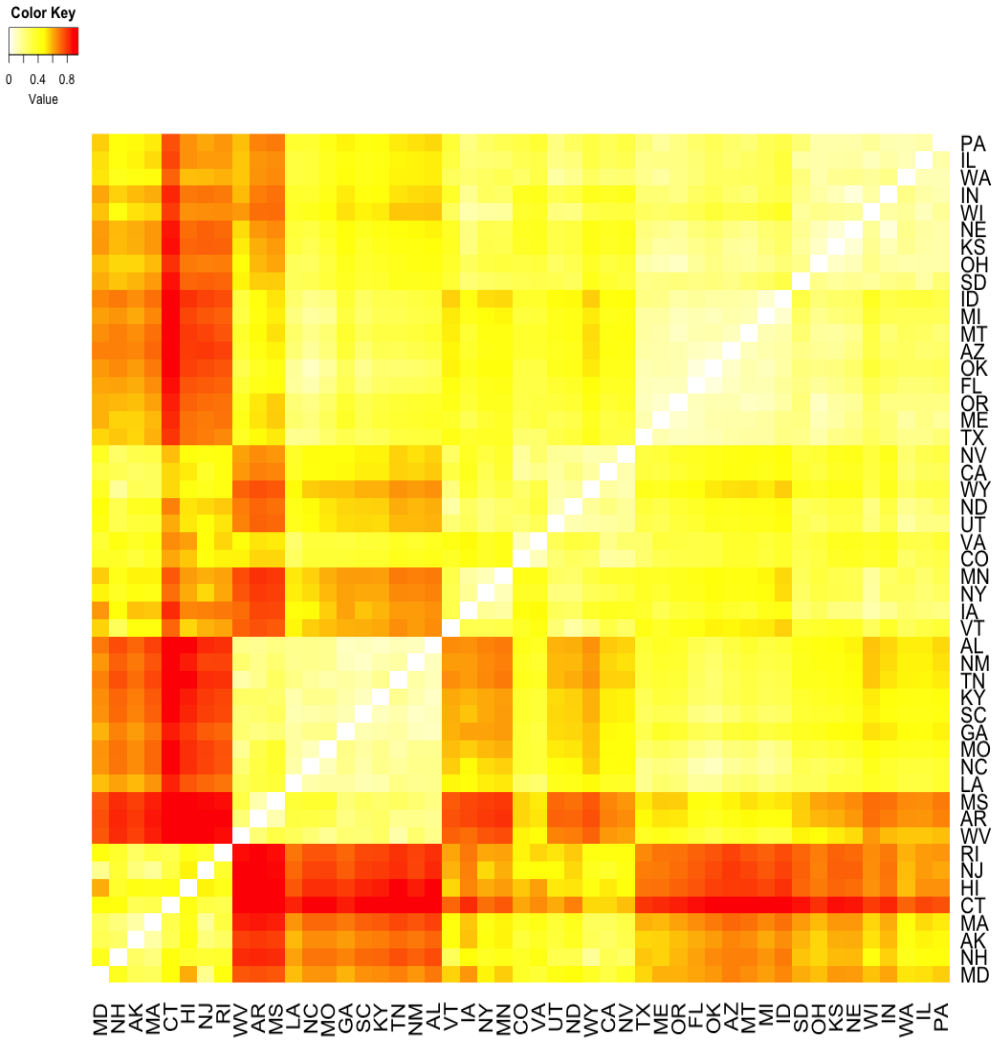


Figure 7.8: Total variance distances between states based on the distributions of median household income are visualized using a heatmap.

7.2.3 Results

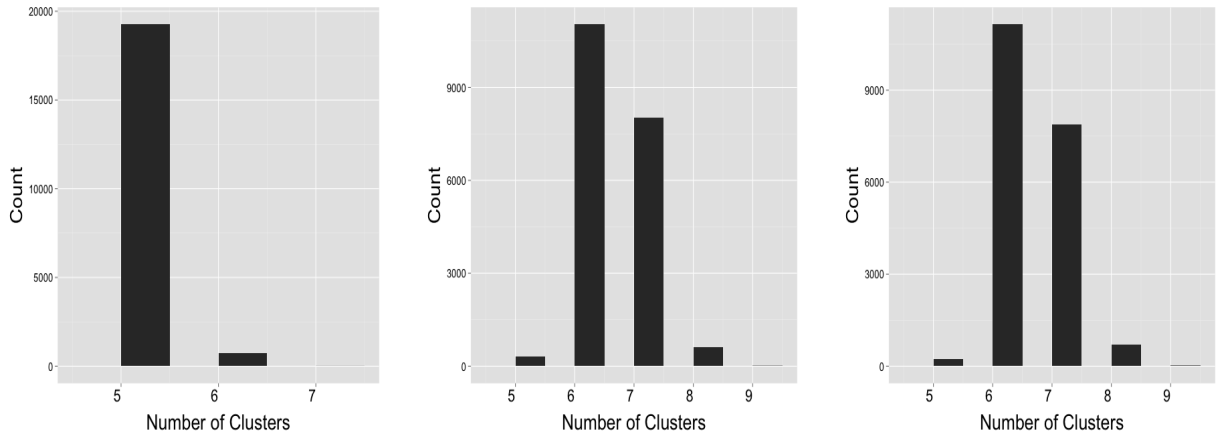
In this section, we present our clustering results inferred by NDP, NDP-MBJ, and LB-NDP. The results provided in this section are based on $20,000$ samples obtained after a $50,000$ burn-in period. We provide the posterior probabilities of the number of distinct states obtained by three models in Table 7.1. Table 7.1 demonstrates that NDP is strongly in favor of five groups of distinct states with the posterior probability of 0.963. On the other hand, both NDP-MBJ and LB-NDP prefer six groups of distinct states (the posterior probabilities 0.05175 and 0.5572, respectively). We infer the fact that the concentration parameter α in NDP is 1 and the posterior mean of $\hat{\alpha}$ estimated via LB-NDP is 2.64 causes the different number of distinct states. Also, the different numbers of distinct states estimated by NDP, NDP-MBJ, and LB-NDP result in different clustering memberships for 49 states. Based on the posterior probability on the number of distinct states, we decide to accept five partitions of the sets of states for NDP and six partitions of the sets of states for both NDP-MBJ and LB-NDP.

The resulting partitions are listed in Tables 7.2, 7.3, and 7.4 with their labels. As we provide the five partitions of the sets of states for NDP, we leave two cells in the row for C_3 in Table 7.2 for easier comparison with NDP-MBJ and LB-NDP. Additionally, the clustering assignments are visualized in Figure 7.10. Alaska and Hawaii are excluded from the maps in Figure 7.10. Figure 7.10 shows that some partitions of the sets of states inferred from NDP, NDP-MBJ, and LB-NDP are not same but similar. When the three models are applied, all three models show the different clustering memberships for the two states, WY and UT. We

Table 7.1: The posterior probabilities of the number of distinct clusters are provided for three different models. The cell with light gray marks the highest probability at each model.

	The Number of Distinct Clusters					
Model	5	6	7	8	9	10
NDP with $\alpha = 1$ and $\beta = 1$	0.96300	0.03675	0.00025	0	0	0
NDP-MBJ	0.01570	0.55175	0.40130	0.03080	0.00045	0
LB-NDP with TVD	0.0119	0.5572	0.3947	0.0354	0.0008	0

may infer WY and UT are the states that share the similar characteristics to other states in different clusters. For example, NDP assigns WY and UT to the same partition with IL, IN, KS, MN, NE, NY, ND, OH, PA, VT, WA, and WI. However, LB-NDP segments WY into the partition (C_1 in Table 7.4) that is consisted of the following states: CT, MD, MA, NH, NJ, RI, AK, and HI. Figure 7.11 depicts the median household income from counties or county equivalents by states based on the clustering solutions obtained by the three models. LB-NDP assigns WY into the cluster that includes AK, MD, NH, and etc. Figure 7.11(c) shows that the distribution of the median household income in WY looks similar to the distribution of the median household income in NH in that WY and NH have the high density of the counties near $\$60,000$ and have some counties clustered near $\$75,000$. Also, LB-NDP assigns UT into the cluster that includes CA because the distribution of the median household income in CA covers up the distribution of the median household income in UT. Also, the distribution in UT have the similar features with the distribution of the median household income in ND in that the median household incomes of most counties in both states are peaked at near $\$50,000$. However, NDP-MBJ assigns WY into the cluster that includes CA, ND, and NV. NDP model assigns WY into the cluster that includes IL, IN, and etc. We may conclude that WY is the state that shares the similarity and neighbors with



(a) NDP with $\alpha = 1$ and $\beta = 1$ (b) NDP-MBJ (c) LB-NDP with TVD

Figure 7.9: The number of distinct states estimated by NDP, NDP-MBJ, and LB-NDP.

the following clusters: C_4 in Figure 7.11(a) C_3 in Figure 7.11(b), and C_1 in Figure 7.11(c). Also, UT is that state that neighbors with C_4 in Figure 7.11(a), C_5 in Figure 7.11(b), and C_3 in Figure 7.11(c). However, the three models show that both UT and WY are distinct from the states in C_4 and C_6 in Figure 7.11(c).

Interestingly, the two states, VA and CO, are clustered in the same group. Figure 7.11 shows that the distributions of the median household income in VA and CO have the larger dispersion than other states. The shape of the distributions of the median household income in VA and CO looks similar because of the high density of the median household income that ranges from approximately $\$30,000$ to $\$70,000$. Also, there are the sub-cluster of counties with the median household income higher than $\$100,000$.

We also examine the clustering structure of the counties within states. However, the interpretation of the sub-clusters should be conditional to the partition of the sets of states via NDP, NDP-MBJ, or LB-NDP. The numbers of sub-clusters in the partitions of the states are provided in Tables 7.2, 7.3, and 7.4. From the linkage based Dirichlet process mixture model, we obtained six groups of clustered states and labeled them as C_1, C_2, C_3, C_4, C_5, and C_6. Each group of clustered states had 2, 3, 3, 2, 2, and 3 sub-clusters, respectively. We have identified that the states that are categorized into Cluster C_6 in Tables 7.2, 7.3, and 7.4 are same for the three models. However, the numbers of sub-clusters for $C-6$ are different; 4, 5, and 3 for NDP, NDP-MBJ, and LB-NDP, respectively. We have calculated the silhouette coefficients in Table 7.5 for the partitions in Tables 7.2, 7.3, and 7.4. It shows that LB-NDP provides the highest silhouette coefficient for C_6 among the three models. We may conclude that three sub-clusters defined by LB-NDP is the optimal clustering solution. Furthermore, LB-NDP finds a better clustering solution than NDP-MBJ and NDP for Cluster C_4 that are consisted of the same states by the three models, but has a different number of sub-clusters.

As the number of distinct states are different for the NDP and LB-NDP models, an ex-

act comparison between the two models with respect to the quality of clustering is not easy. Therefore, we calculate the overall silhouette coefficient for evaluating the performance of the models. In order to obtain the overall silhouette coefficient, we calculate the average of the silhouette coefficient for each partition of clustered states. We then take the average of the averages of the silhouette coefficients from all partitions of clustered states, and define this as the overall silhouette coefficient. The detailed silhouette coefficients for each partition of the states are provided in Table 7.5. Note that we are not able to obtain the silhouette coefficient for cluster C_1 because of the number of sub-clusters within C_1 as shown in Table 7.2 and Table 7.3. The obtained overall silhouette coefficients are 0.517, 0.457, and 0.436 for LB-NDP, NDP-MBJ, and NDP respectively. As the number of the counties within the clustered states can be differ, the estimation of the vector of the concentration parameters $\boldsymbol{\beta}$ is important to obtain the optimal clustering solution. We conclude that the linkage based nested Dirichlet process mixture model finds a better clustering solution than NDP-MBJ and the nested Dirichlet process mixture model when the concentration parameters are fixed at 1.

Table 7.2: The partitions of the distinct states defined by NDP with $\alpha = 1$ and $\beta = 1$ are provided with the number of sub-clusters within the partitions.

Cluster	**State**	**# of sub-clusters**
C_1	CT, MD, MA, NH, NJ, RI, AK, HI	1
C_2	CA, CO, NV, VA	4
C_3	-	-
C_4	AZ, FL, ID, LA, ME, MI, MO, MT, NC, OK, OR, SD, TX	3
C_5	IL, IN, IA, KS, MN, NE, NY, ND, OH, PA, UT, VT, WA, WI, WY	2
C_6	AL, AR, GA, KY, MS, NM, SC, TN, WV	4

Table 7.3: The partitions of the distinct states defined by NDP-MBJ are provided with the number of sub-clusters within the partitions.

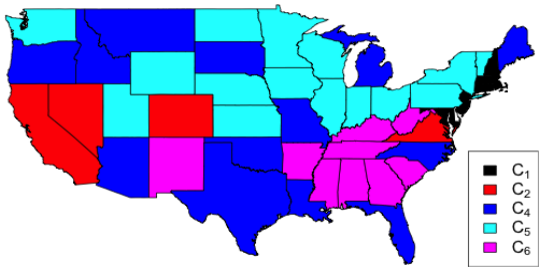
Cluster	**State**	**# of sub-clusters**
C_1	CT, MD, MA, NH, NJ, RI, AK, HI	1
C_2	CO, VA	3
C_3	CA, NV, ND, WY	2
C_4	AZ, FL, ID, LA, ME, MI, MO, MT, NC, OK, OR, SD, TX	3
C_5	IL, IN, IA, KS, MN, NE, NY, OH, PA, UT, VT, WA, WI	3
C_6	AL, AR, GA, KY, MS, NM, SC, TN, WV	5

Table 7.4: The partitions of the distinct states defined by the linkage based nested Dirichlet process are provided with the number of sub-clusters within the partitions.

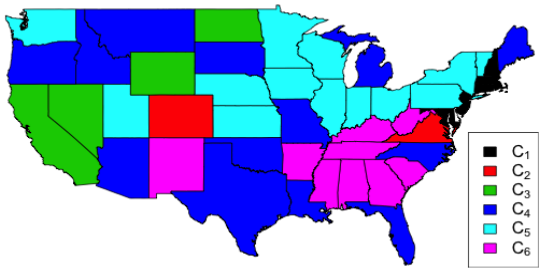
Cluster	**State**	**# of sub-clusters**
C_1	CT, MD, MA, NH, NJ, RI, WY, AK, HI	2
C_2	CO, VA	3
C_3	CA, NV, ND, UT	3
C_4	AZ, FL, ID, LA, ME, MI, MO, MT, NC, OK, OR, SD, TX	2
C_5	IL, IN, IA, KS, MN, NE, NY, OH, PA, VT, WA, WI	2
C_6	AL, AR, GA, KY, MS, NM, SC, TN, WV	3

Table 7.5: Silhouette coefficients obtained by the linkage based nested Dirichlet process model and the nested Dirichlet process mixture model are provided for clustered states.

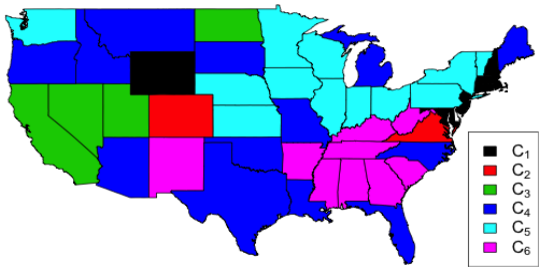
	Cluster Label					
Model	C_1	C_2	C_3	C_4	C_5	C_6
NDP with $\alpha = \beta = 1$	NA	0.208	–	0.428	0.664	0.442
NDP-MBJ	NA	0.611	0.642	0.429	0.187	0.417
LBNDP with TVD	0.343	0.630	0.636	0.621	0.231	0.643



(a) NDP with $\alpha = 1$ and $\beta = 1$

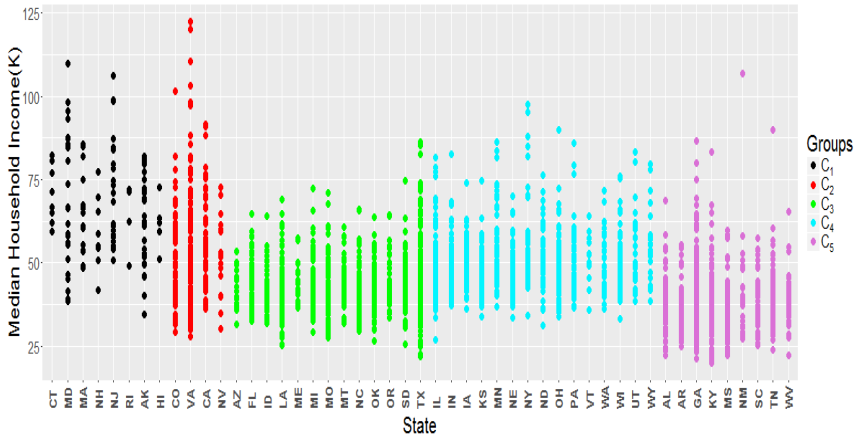


(b) NDP-MBJ

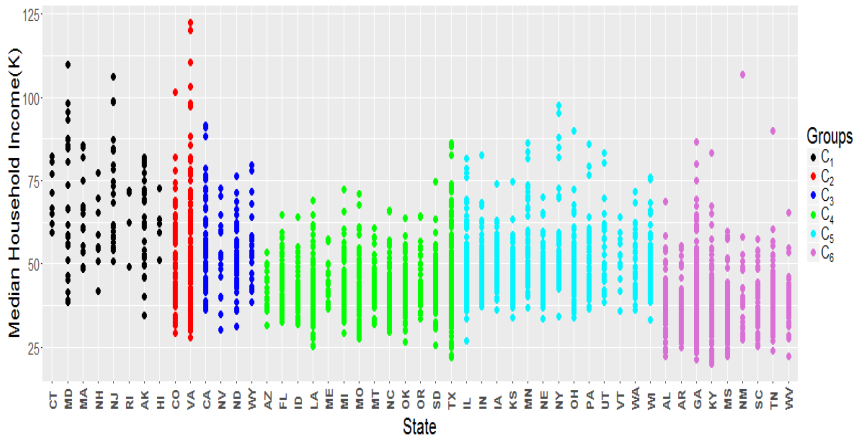


(c) LB-NDP with the total variation distance

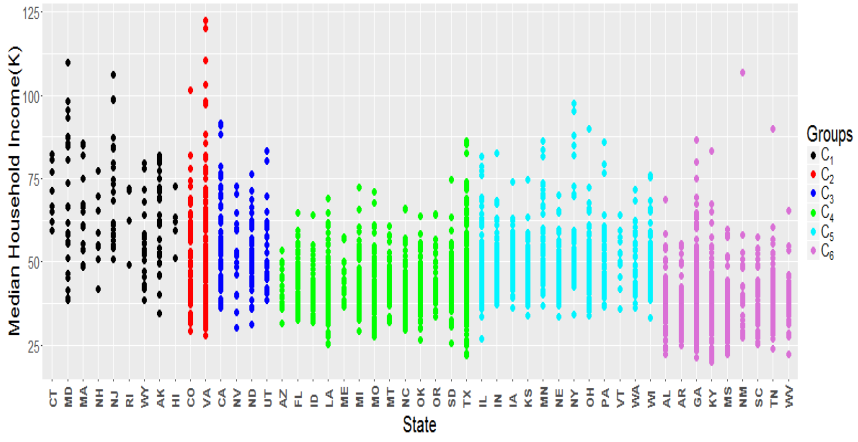
Figure 7.10: States are visualized on the map, and each state is colored based on its partition defined by the model. Panels (a), (b), and (c) provide the clustering solutions by NDP with $\alpha = \beta = 1$, NDP-MBJ, and LB-NDP with the total variation distance, respectively.



(a) NDP



(b) NDP-MBJ



(c) LB-NDP

Figure 7.11: The median household income from counties or county equivalents are plotted by states based on the clustering solution.

7.3 Conclusion

The linkage based nested Dirichlet process estimates the concentration parameter α for clustering distinct densities (i.e. states), and infers the vector of the concentration parameter $\boldsymbol{\beta}$ for obtaining sub-clusters (i.e. counties) within the partitions of the distinct densities. This fact can provide users with more flexible modeling results with respect to identifying sub-clusters and densities simultaneously. In Chapter 7, through a sequence of the simulation studies, we have examined the performance of the linkage based nested Dirichlet process mixture model. As a conclusion, the linkage based nested Dirichlet process mixture model demonstrates better performance than the nested Dirichlet process mixture model with NDP-MBJ, given a distribution of a small number of observations. Additionally, when the distributions are overlapping, the linkage based nested Dirichlet process mixture model shows a better performance than the nested Dirichlet process mixture model with NDP-MBJ with respect to the overall silhouette coefficients.

As an application of the linkage based nested Dirichlet process, we have modeled the median household income data using the linkage based nested Dirichlet process mixture model, the nested Dirichlet process mixture model with NDP-MBJ, and the nested Dirichlet process mixture model. We have shown that the linkage based nested Dirichlet process has performed better among the three models based on the calculation of the silhouette coefficient. In this chapter, we conclude that the estimation of concentration parameters for multi-level clustering is important because the choice of the concentration parameters can affect the clustering solutions.

Chapter 8

Discussion and Future Work

In this work, we have introduced the two methods, the linkage based Dirichlet process and the linkage based nested Dirichlet process, for estimating the concentration parameter in the DP and the concentration parameters in the nested DP. Other studies for estimating the concentration parameter have used only the defined number of clusters and the number of data points. However, our techniques calibrate the probability distances between the clusters and the configuration of the clusters, so that we maximize information based on the observations in the defined clusters and use it as much as possible. We have shown that two methods are capable of providing the optimal clustering solutions based on the calculation of the silhouette coefficient through the sequence of the simulation studies and applications

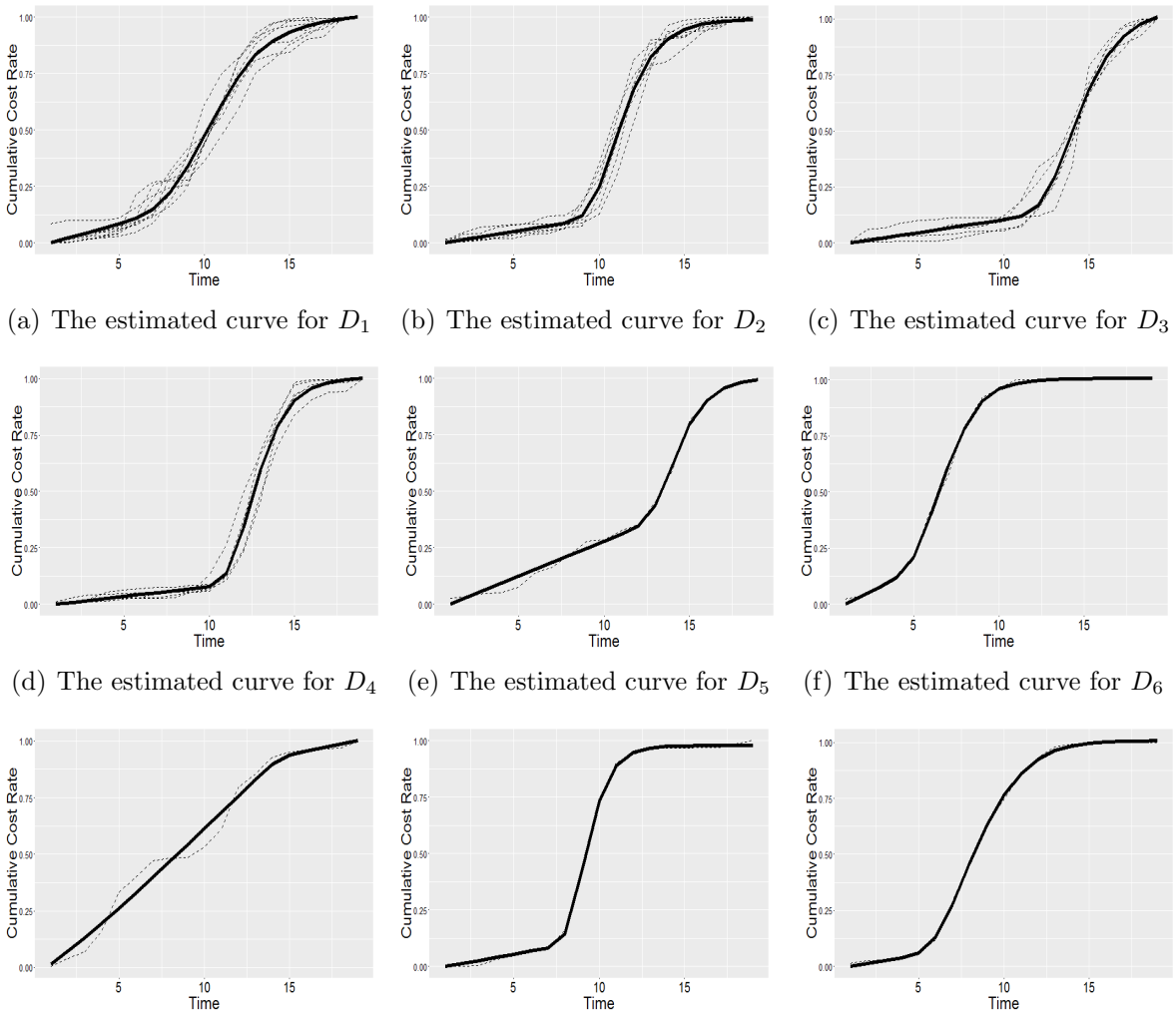
The linkage based Dirichlet process mixture model and the linkage based nested Dirichlet process mixture model have demonstrated themselves to be a useful statistical tool as the clustering algorithms. Currently, the dimension of the simulated data for the linkage based nested Dirichlet process mixture model is up to 2 due to the difficulty in measuring the distance between the mixtures of the distributions. By using other measures introduced in Nowakowska et al. (2014), we may extend our simulation study for the linkage based nested Dirichlet process to the higher dimensional data.

In this work, the linkage based Dirichlet process and the linkage based nested Dirichlet process are used for clustering analyses. However, the linkage based Dirichlet process and the linkage based nested Dirichlet process are not limited to clustering analyses. Like the Dirichlet process, these two techniques also can be used in random effect models as a prior for a random effect term. For future, we may perform random effect models in linear regressions by using the linkage based Dirichlet process or the linkage based nested Dirichlet process.

Through the sequence of simulation studies for a small number of observations and modeling the timeline for the building construction costs, we have examined the effect of the estimation of the concentration parameter, and we have shown that the linkage based Dirichlet process provides a better clustering solution. Recently, clustering analyses for a small number of observations from a high dimensional space become popular in many areas such as modern biology; i.e. gene expression clustering. Thus, we may apply our techniques for clustering gene expression data.

Chapter 9

Appendix



(a) The estimated curve for D_1 (b) The estimated curve for D_2 (c) The estimated curve for D_3

(d) The estimated curve for D_4 (e) The estimated curve for D_5 (f) The estimated curve for D_6

(g) The estimated curve for D_7 (h) The estimated curve for D_8 (i) The estimated curve for D_9

Figure A.1: 9 estimated curves by DP-MBJ are illustrated with their members.

Chapter 10

Bibliography

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723.

Aldous, D. J. (1985). *Exchangeability and related topics*. Springer.

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian non-parametric problems. *The annals of statistics*, pages 1152–1174.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

Balakrishnan, N. (2001). *Continuous multivariate distributions*. Wiley Online Library.

Berger, R. (1981). Comparison of the gompertz and logistic equations to describe plant disease progress. *Phytopathology*, 71(7):716–719.

Bhattacharyya, A. (1946). On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics*, pages 401–406.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(7):719–725.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via pólya urn schemes. *The annals of statistics*, pages 353–355.

Blasco, A., Piles, M., Varona, L., et al. (2003). A bayesian analysis of the effect of selection for growth rate on growth curves in rabbits. *Genetics Selection Evolution*, 35(1):21–42.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Brent, R. P. (1973). *Algorithms for minimization without derivatives*. Courier Dover Publications.

Casella, G. (1992). Illustrating empirical bayes methods. *Chemometrics and intelligent laboratory systems*, 16(2):107–125.

Casella, G. (2001). Empirical bayes gibbs sampling. *Biostatistics*, 2(4):485–500.

Chen, H., Chung, W., Xu, J. J., Wang, G., Qin, Y., and Chau, M. (2004). Crime data mining: a general framework and some examples. *Computer*, 37(4):50–56.

Chen, S. S. and Gopalakrishnan, P. S. (1998). Clustering via the bayesian information criterion with applications in speech recognition. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 2, pages 645–648. IEEE.

Chung, J., Kannappan, P., Ng, C., and Sahoo, P. (1989). Measures of distance between probability distributions. *Journal of mathematical analysis and applications*, 138(1):280–292.

Clarke, C. R., Chinchilla, D., Hind, S. R., Taguchi, F., Miki, R., Ichinose, Y., Martin, G. B., Felix, G., and Vinatzer, B. A. (2013). Allelic variation in two distinct pseudomonas syringae flagellin epitopes modulates the strength of plant immune responses but not bacterial motility. *New Phytologist*, 200(3):847–860.

Clough, R. H., Sears, G. A., and Sears, S. K. (2000). *Construction project management*. John Wiley & Sons.

Dagan, I., Lee, L., and Pereira, F. (1997). Similarity-based methods for word sense disambiguation. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 56–63. Association for Computational Linguistics.

Dahl, D. B. (2006). Model-based clustering for expression data via a dirichlet process mixture model. *Bayesian inference for gene expression and proteomics*, pages 201–218.

DeNavas-Walt, C. (2010). *Income, poverty, and health insurance coverage in the United States (2005)*. DIANE Publishing.

Dhillon, I. S. and Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine learning*, 42(1-2):143–175.

Dorazio, R. M. (2009). On selecting a prior for the precision parameter of dirichlet process mixture models. *Journal of Statistical Planning and Inference*, 139(9):3384–3390.

Dunford, N. and Schwartz, J. T. (1958). Linear operators, vol. i. *Interscience, New York*, 1963.

Escobar, M. D. (1994). Estimating normal means with a dirichlet process prior. *Journal of the American Statistical Association*, 89(425):268–277.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588.

Fabius, J. (1973). Two characterizations of the dirichlet distribution. *The Annals of Statistics*, pages 583–587.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

Figueiredo, M. A. and Jain, A. K. (2002). Unsupervised learning of finite mixture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(3):381–396.

Fisk, E. R. (1988). Construction project adminstration.

Fraley, C. and Raftery, A. E. (1998). How many clusters? which clustering method? answers via model-based cluster analysis. *The computer journal*, 41(8):578–588.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631.

Gaffney, S. J. and Smyth, P. (2003). Curve clustering with random effects regression mixtures. In *Proceedings of the ninth international workshop on artificial intelligence and statistics*. Citeseer.

Gibbs, A. L. and Su, F. E. (2002). On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435.

Goldstein, H. (1965). *Classical mechanics*. Pearson Education India.

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical transactions of the Royal Society of London*, 115:513–583.

Görür, D. (2007). *Nonparametric Bayesian discrete latent variable models for unsupervised learning*. PhD thesis, Berlin Institute of Technology.

Gould, F. E. and Joyce, N. E. (2003). *Construction project management*. Prentice Hall Upper Saddle River, NJ.

Hardle, W. (1990). *Applied nonparametric regression*, volume 27. Cambridge Univ Press.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Applied statistics*, pages 100–108.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., and Tibshirani, R. (2009). *The elements of statistical learning*, volume 2. Springer.

Heard, N. A., Holmes, C. C., and Stephens, D. A. (2006). A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of bayesian hierarchical clustering of curves. *Journal of the American Statistical Association*, 101(473):18–29.

Hellinger, E. (1909). Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 136:210–271.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453).

Kaufman, L. and Rousseeuw, P. (1987). *Clustering by means of medoids*. North-Holland.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.

Kim, J. G., Menzefricke, U., and Feinberg, F. M. (2004). Assessing heterogeneity in discrete choice models using a dirichlet process prior. *Review of marketing Science*, 2(1).

Kleinman, K. P. and Ibrahim, J. G. (1998). A semiparametric bayesian approach to the random effects model. *Biometrics*, pages 921–938.

Kolb, W. M. (1983). *Curve fitting for programmable calculators*. Imtec.

Kullback, S. (1987). The kullback-leibler distance.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, pages 79–86.

Lancaster, P. and Salkauskas, K. (1986). *Curve and surface fitting*. Academic press.

Levina, E. and Bickel, P. (2001). The earth mover's distance is the mallows distance: some insights from statistics. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 251–256. IEEE.

Linoff, G. S. and Berry, M. J. (2011). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons.

Liu, J. S. (1996). Nonparametric hierarchical bayes via sequential imputations. *The Annals of Statistics*, pages 911–930.

MacEachern, S. N. (1994). Estimating normal means with a conjugate style dirichlet process prior. *Communications in Statistics-Simulation and Computation*, 23(3):727–741.

MacEachern, S. N. (2000). Dependent dirichlet processes. *Unpublished manuscript, Department of Statistics, The Ohio State University*.

MacEachern, S. N. and Müller, P. (1998). Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238.

McAuliffe, J. D., Blei, D. M., and Jordan, M. I. (2006). Nonparametric empirical bayes for the dirichlet process mixture model. *Statistics and Computing*, 16(1):5–14.

McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture. *Applied Statistics*, pages 318–324.

McLachlan, G. J. and Basford, K. E. (1988). Mixture models. inference and applications to clustering. *Statistics: Textbooks and Monographs, New York: Dekker, 1988*, 1.

Medvedovic, M. and Sivaganesan, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, 18(9):1194–1206.

Minka, T. (2000). Estimating a dirichlet distribution.

Motulsky, H. and Christopoulos, A. (2004). *Fitting models to biological data using linear and nonlinear regression: a practical guide to curve fitting*. Oxford University Press.

Motulsky, H. J. and Ransnas, L. A. (1987). Fitting curves to data using nonlinear regression: a practical and nonmathematical review. *The FASEB journal*, 1(5):365–374.

Mukhopadhyay, S. and Gelfand, A. E. (1997). Dirichlet process mixed generalized linear models. *journal of the American Statistical Association*, 92(438):633–639.

Müller, P., Erkanli, A., and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 83(1):67–79.

Murray, A. T., McGuffog, I., Western, J. S., and Mullins, P. (2001). Exploratory spatial data analysis techniques for examining urban crime implications for evaluating treatment. *British Journal of Criminology*, 41(2):309–329.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142.

Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265.

Nikulin, M. S. (2001). Hellinger distance. *Encyclopeadia of Mathematics, edited by Hazewinkel, Michiel, CUP, Springer, http://www. encyclopediaofmath. org/index. php.*

Nowakowska, E., Koronacki, J., and Lipovetsky, S. (2014). Tractable measure of component overlap for gaussian mixture models. *arXiv preprint arXiv:1407.7172*.

Nsoesie, E. O., Leman, S. C., and Marathe, M. V. (2014). A dirichlet process model for classifying and forecasting epidemic curves. *BMC infectious diseases*, 14(1):12.

Pitman, J. et al. (2002). Combinatorial stochastic processes. Technical report, Technical Report 621, Dept. Statistics, UC Berkeley, 2002. Lecture notes for St. Flour course.

Pyle, D. (1999). *Data preparation for data mining*, volume 1. Morgan Kaufmann.

Rabaoui, A., Viandier, N., Marais, J., and Duflos, E. (2011). On selecting the hyperparameters of the dpm models for the density estimation of observation errors. In *Acoustics,*

Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pages 4092–4095. IEEE.

Ramsay, J. O. (2006). *Functional data analysis*. Wiley Online Library.

Rasmussen, C. E. (1999). The infinite gaussian mixture model. In *NIPS*, volume 12, pages 554–560.

Ray, S. and Turi, R. H. (1999). Determination of number of clusters in k-means clustering and application in colour image segmentation. In *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques*, pages 137–143.

Reutter, A. and Johnson, V. E. (1995). *General strategies for assessing convergence of MCMC algorithms using coupled sample paths*. Citeseer.

Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2008). The nested dirichlet process. *Journal of the American Statistical Association*, 103(483).

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Sethuraman, J. (1991). A constructive definition of dirichlet priors. Technical report, DTIC Document.

Shahbaba, B. and Neal, R. (2009). Nonlinear models using dirichlet process mixtures. *The Journal of Machine Learning Research*, 10:1829–1850.

Silverman, B. and Ramsay, J. (2005). *Functional Data Analysis*. Springer.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–52.

Steinbach, M., Karypis, G., Kumar, V., et al. (2000). A comparison of document clustering techniques. In *KDD workshop on text mining*, volume 400, pages 525–526. Boston, MA.

Susarla, V. and Van Ryzin, J. (1976). Nonparametric bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association*, 71(356):897–902.

Tarpey, T. (2007). Linear transformations and the k-means clustering algorithm. *The American Statistician*, 61(1).

Teh, Y. W. (2010). Dirichlet process. In *Encyclopedia of machine learning*, pages 280–287. Springer.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476).

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Trauner, T. J. (1993). *Managing the construction project: a practical guide for the project manager*, volume 2. John Wiley & Sons Incorporated.

Wakita, K. and Tsurumi, T. (2007). Finding community structure in mega-scale social networks:[extended abstract]. In *Proceedings of the 16th international conference on World Wide Web*, pages 1275–1276. ACM.

West, M. (1992). *Hyperparameter estimation in Dirichlet process mixture models*. Duke University.

West, M. and Escobar, M. D. (1993). *Hierarchical priors and mixture models, with application in regression and density estimation*. Institute of Statistics and Decision Sciences, Duke University.

Zhou, B. and Hansen, J. H. (2000). Unsupervised audio stream segmentation and clustering via the bayesian information criterion. In *INTERSPEECH*, pages 714–717.