ABSTRACT

KANG, CHANGKU. Regression via clustering using Dirichlet mixtures. (Under the direction of Subhashis Ghosal.)

Regression analysis is a fundamental problem of statistics. When the regression function has an unknown form, parametric analysis is sometimes inappropriate. In such a situation, the regression function should be estimated by nonparametric methods. Often, the regressor variable is sampled from several different subpopulations and the regression function has different forms depending on the source. The labels of these source subpopulations are not observable. Although a nonparametrically specified regression function can capture the overall regression function, nonparametric regression estimates are usually dependent on the assumption of homoscedasticity of additive errors. If the underlying distribution of X has unknown clusters, then the usual assumption, the homoscedasity does not hold. In estimating the regression function, we propose the idea of first finding clusters in the regressor variables by the Dirichlet mixture to impute lost subpopulation labels. A standard regression method such as linear or polynomial regression then may be used within each cluster. Markov Chain Monte Carlo (MCMC) sampling method is used to find the clusters and for each sample the estimated regression functions can be obtained. We also apply our method to the large p, small n problem, where the number of variables p is much greater than the number of samples n. In several simulation experiments, our method is compared to other methods such as kernel and smoothing splines in the univariate case and GAM (generalized additive model) and MARS (Multivariate Adaptive Regression Splines) in the multivariate case. The consistency issue is

discussed without explicit proof.

Regression via clustering using Dirichlet mixtures

by

Changku Kang

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh

2005

Submitted on : 10/06/2005

APPROVED BY:

Subhashis Ghosal Chair of Advisory Committee	John F. Monahan
Sujit K. Ghosh	Hao H. Zhang

To my family

Biography

Changku Kang was born in Naju, the Republic of Korea on August 5, 1972. He entered Department of Mathematics at Hanyang University in 1992 and took his B.S. in Mathematics in 1999. He obtained his M.S. in Statistics in May, 2001 at Hanyang University. He joined Department of Statistics, North Carolina State University for the pursuit of Ph.D. degree in August, 2001.

Acknowledgements

I would like to express my deepest gratitude and biggest appreciation to my advisor Dr. Subhashis Ghosal. The members of my dissertation committee, Dr. John F. Monahan, Dr. Sujit K. Ghosh and Dr. Hao H. Zhang provided me with extremely useful information and valuable advice. I sincerely thank all of them.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

A common problem of data analysis in modern statistics is the estimation of a regression function based on sampled data $(X_1, Y_1), \ldots, (X_n, Y_n)$. The regression function, that is the mean of Y given X, is often specified as a given function such as a polynomial of certain order with unknown coefficients. Suppose the underlying distribution of X is a mixture of several distributions in that X may arise from different subpopulations. In such a situation, it is quite plausible that the regression function may have different parameters when the sample comes from different groups. However, we do not observe the group labels. The mixture distribution can arise in many real situations such as it does in biology or economics. For example, in marketing data, suppose that consumers rate the quality of a product. Different customers may give different weights to various factors depending upon their background and mentality. In other words, the population actually consists of different

subpopulations where different regression functions are in effect, but subpopulation membership is abstract and is not observable. Had we observed the labels, regression analysis would have been straightforward. In the absence of the labels, we use the auxiliary measurements to impute the labels and use simple regression analysis within each hypothetical group. It is not unreasonable to expect that subjects within similar measurement are likely to have similar background and mentality. If customers of different types are identified, we may use simple regression analysis within each cluster and the true overall regression function will be estimated as a weighted combination of all regression estimates. In the whole analysis, uncertainty in group membership as well as the number of groups should be taken into consideration.

In this thesis we propose a method to estimate an unknown regression function by splitting it into an unknown number of clusters and then using some simple regression models within each cluster. In finding the clusters, a Bayesian nonparametric method is considered. Standard regression methods are used in fitting a regression function within each cluster. We also give an intuitive argument why consistency of our estimator is expected.

Suppose we are given data $(X_1, Y_1), \ldots, (X_n, Y_n)$ where X_i is p-variate continuous variables and Y_i is univariate. We want to consider the regression model,

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where $\varepsilon_i \overset{iid}{\sim} N(0, T^2)$. It is not assumed that $f(\cdot)$ has a specific structure, such as linearity. Now, let the population of X consist of several groups, say k groups, which

are, respectively, distributed as

$$N(\mu_1, \Sigma_1), \ldots, N(\mu_k, \Sigma_k); \tag{1.2}$$

here μ_j is the mean vector and Σ_j is the $p \times p$ variance covariance matrix corresponding to the j^{th} group. Let J be the latent variable which indicates the index of the group where the observation belongs. The prior distribution of J is given by $\pi_j = P(J = j)$ and thus the posterior distribution given that an X observation is equal to x is given by $\pi_j(x) = P(J = j|X = x)$. Let the conditional distribution of X given J and the conditional distribution of Y given X and J be given by

$$X|J = j \sim N(\mu_j, \Sigma_j), \tag{1.3}$$

$$Y|X = x, J = j \sim N(f_j(x), T_j^2), \quad j = 1, \ldots, k,$$

where $f_j(x)$ is the regression function in group j. Then, by the law of total probabilities,

$$E(Y|X = x) = \sum_{j=1}^{k} E(Y|X = x, J = j) \cdot P(J = j|X = x) = \sum_{j=1}^{k} f_j(x)\pi_j(x). \tag{1.4}$$

Also, the conditional variance is given by

$$\begin{aligned} \mathrm{Var}(Y|X = x) &= E_J(\mathrm{Var}(Y|X = x, J)) + \mathrm{Var}_J(E(Y|X = x, J)) \\ &= E_J(T_J^2) + \mathrm{Var}_J(f_J(x)) \\ &= \sum_{j=1}^{k} T_j^2\pi_j + \sum_{j=1}^{k} f_j^2(x)\pi_j(x) - \Big(\sum_{j=1}^{k} f_j(x)\pi_j(x)\Big)^2. \end{aligned} \tag{1.5}$$

Note here that if T_j^2 are all equal, then first term in (1.5) is constant but the other terms may not be constant in x. For example, Figure 1.1 shows that there are two groups and the conditional mean and variance are given. That is, the regression

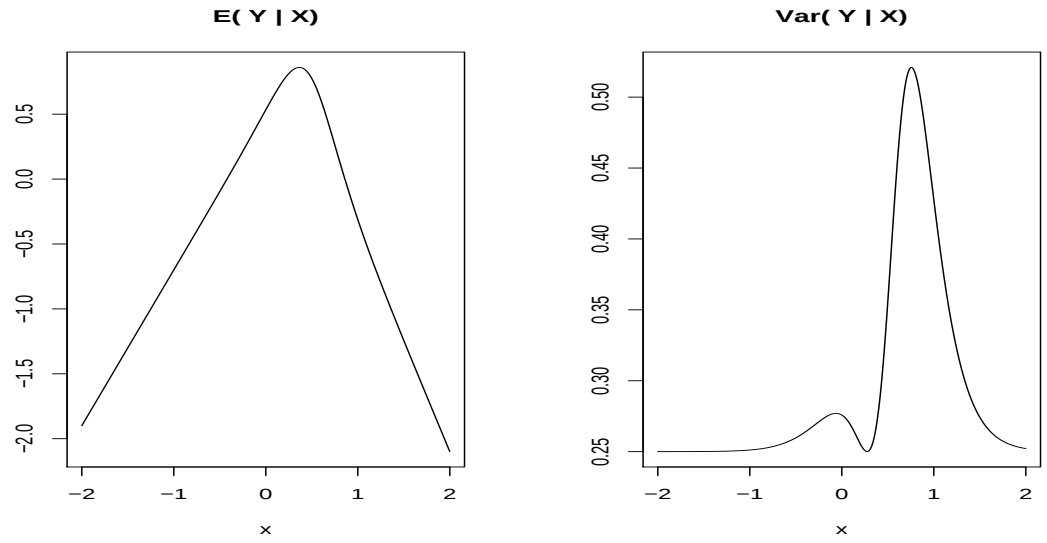


Figure 1.1: The plot of the conditional expectation and variance which depend on the value of X. X is distributed by the mixture of two normals with $\pi_1 = \pi_2 = 0.5$, $\mu_1 = 0, \mu_2 = 1$, $\Sigma_1 = \Sigma_2 = 0.4^2$, $f_1(x) = 0.5 + 1.2x, f_2(x) = 1.3 - 1.7x$, and $T_1^2 = T_2^2 = 0.5^2$.

may not be homoscedastic. This shows that usual nonparametric regression model is unable to handle this simple and intuitive structure in model (1.3). Of course heteroscedastic nonparametric model contains (1.3), but it is very difficult to estimate the regression function under that scenario.

The idea of estimation of the regression function in this situation has two steps. In the first step, as we assume that the underlying distribution of X follows the mixture distribution, we may try to recover lost subpopulation labels by identifying clusters in the data. Second, once the clusters have been identified, standard parametric regression techniques may be used within each cluster.

To form the clusters, one of most popular methods in Bayesian nonparametrics, namely that of Dirichlet mixture (DM), will be used (see Chapter 3). We assign

Dirichlet mixture of normal prior for the density of X observations which means that we model the density of X as a mixture of normal densities and true mixing distribution given the Dirichlet process prior. Because Dirichlet samples are discrete distributions and observations arise according to a Polya urn scheme (see Chapter 3), the Dirichlet mixture prior has the ability to automatically produce clusters. The means of the hidden groups are generated by an MCMC algorithm. In fact, we only need the configuration of the ties among the latent means to identify the clusters at each MCMC step but the group means and hyperparameters need not be updated within MCMC iteration. This will allow substantial reduction of computational complexity at the MCMC step.

Once the clusters have been identified, we can use a standard regression method, such as the method of least squares for linear regression, within each cluster. Polynomial regression may be used if we suspect the lack of linearity in the model. In the multidimensional case, a parsimonious model building is an important issue so that a variable selection method is usually used in regression problems. A useful method which can automatically select the important variables while estimating regression is given by the LASSO regression method [Tibshirani (1996)]. Unlike other shrinkage methods such as the ridge regression [Hoerl and Kennard (1988)], LASSO can shrink certain coefficients to exactly zero depending on the value of some tuning parameter. A cross validation method can be applied to find the optimal value of the tuning parameter.

When we use the Dirichlet mixture model, we do not do a fully Bayesian analysis because the computation time of full MCMC updating procedure may be too long.

For instance, the inversion of high dimensional covariance matrix may be needed for the fully Bayesian method. Instead we use simple non-Bayesian estimates so that we can save substantial computing time.

There is a challenging problem, so-called the large p small n problem, where the number of variables p is large relative to the number of samples n. This happens in many applications such as the microarray data analysis. The method of least squares is inappropriate to handle this problem because of singularity while LASSO tends to have erratic behavior. We propose using the elastic net method suggested by Zou and Hastie (2005) and use it in each cluster to estimate parameters of each regression function after DM procedure finds the clusters.

We will compare our method with several nonparametric methods. A lot of methods using local smoothing such as the kernel method and spline smoothing in the univariate case were provided by many authors. These methods can be applied when there is only one regressor variable. In the multivariate case, the data are sparsely distributed even for a large sample size. This well known difficulty is called "the curse of dimensionality". Many standard nonparametric techniques suffer from this problem. Therefore, it is necessary to consider certain approaches to overcome this kind of problem, especially in a high dimensional situation. Among the important nonparametric regression techniques in higher dimension are Multivariate Adaptive Regression Splines (MARS) [Friedman (1991)] and Generalized Additive Model (GAM) [Hastie and Tibshirani (1990)]. These methods are briefly described in Chapter 2.

In comparison to other methods, we conduct several simulation experiments and

evaluate L_1-error and L_2-error given in (5.1). These are common measures of distance between true regression function and the estimated function in a global sense. An advantage of our method is that there is no need to select to the bandwidth parameter. In the Dirichlet mixture model in (3.1), we need to specify constant M, distribution G_0 and prior for σ. The parameter M is called "precision parameter" because of its role in controlling the spread of the Dirichlet process. We can specify the prior for M but a fixed small value is used here. The posterior distribution of k only depends on the number of clusters in certain MCMC steps so we can easily update and get the information of k. In fact, the value of k obtained from MCMC step is an overestimate of the number of clusters. But, this does not harm much in estimating the regression function; see explanations given in Section 5.3.

In this thesis, we mainly do simulation work and compare the results to those of other methods in the proceeding chapters. The literature reviews for non-Bayesian regression methods, including the linear regression model as a simple case and non-parametric method, are given in Chapter 2. Preliminaries used in this thesis, such as the Dirichlet process, are described in Chapter 3. The main procedure and algorithms are presented in Chapter 4. In Chapter 5, we summarize the simulation results in univariate and multivariate situations, paying special attention to the large p, small n problem. We also analyze a real data collected by a nonprofit organization as an application of our method. Finally some conclusions and future work are given in Chapter 6.

Chapter 2

Non-Bayesian and Bayesian Methods for Regression

In this chapter we review the literature about the regression method from the non-Bayesian and Bayesian nonparametric perspectives. Suppose that there are a vector $X = (X^1, \ldots, X^p)$ and a real valued random variable Y of interest which is influenced by X. Regression describes the approximate relationship between X and Y given data $(X_1, Y_1) \ldots, (X_n, Y_n)$ where $X_i = (X_i^1, \ldots, X_i^p)$. The variable X is called the independent or predictive variable and Y is called the dependent or response variable.

The general heteroscedastic regression model is defined by

$$Y_i = f(X_i) + \varepsilon_i, \tag{2.1}$$

where ε_i is usually assumed by the independent normal distribution having mean 0 and variance $\sigma_i^2(x)$. Note that here, the conditional variance of Y given X depends

on the covariate vector X. If the functional form of $f(\cdot)$ such as linearity is not given, then the model is nonparametric. The simplest and most popular one is the homoscedastic linear regression as a parametric regression by assuming $f(x) = X\beta$ and $\sigma_i^2(x) = \sigma^2$. In our problem, we focus on using linear regression within each cluster after finding several candidate groups.

The ordinary least squares (OLS) method is very popular and a simple method of estimation in linear regression models. We begin with a review of the linear regression along with its modification such as the ridge regression and the LASSO method. Then nonparametric regression method will be reviewed. This method has certain difficulty in the multidimensional case, known as "the curse of dimensionality". Several methods were developed to overcome this drawback. A generalized additive model (GAM) is suggested to restrict the form of the function and multivariate adaptive regression splines (MARS) is developed as the generalization of recursive partitioning regression model.

If we consider Bayesian nonparametric regression methods, we want to use Bayesian method under flexible assumption of the unknown regression function. Gaussian process will be given first. An approach is to consider expansion with respect to some basis for an appropriate function space and put priors on the coefficients of expansion. For examples, the spline basis, the Fourier basis and wavelet bases are commonly used to expand a function. There is an alternative approach proposed by Muller, Erkanli and West (1996) by using density estimation method because the regression function is the conditional expectation under joint distribution of (X, Y).

2.1 Classical Methods

2.1.1 Linear regression

The simplest regression model is the linear regression which assumes that regression function is a linear function of X. Assume that we have the regression model in (1.1). The linear regression model assume $f(\cdot)$ has the linear structure

$$Y_i = \sum_{j=0}^{p} X_{ij}\beta_j + \varepsilon_i, \tag{2.2}$$

where ε_i is distributed by the independent normal distribution having mean 0 and variance σ^2 and $X_{i0} = 1$ for all $i = 1, \ldots, n$. We may assume that X_i's are distributed according to a certain distribution such as a normal distribution. The regression function can be expressed as the conditional expectation of Y given X, $E(Y|X)$. This model includes the polynomial regression model. For example, if X is univariate, then by defining $X_2 = X_1^2$ quadratic relationship can be incorporated into by the model.

Let $\mathbf{X}$ be the $n \times (p+1)$ matrix with $(i,j)^{th}$ component as $X_{i,j}$ and similarly let $\mathbf{Y}$ be the column vector of length n whose i^{th} entry $\mathbf{Y_i}$. Then, model (2.2) can be expressed as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.3}$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$. The OLS estimator is obtained by minimizing the sum of squared error (SSE)

$$\text{SSE } (\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}), \tag{2.4}$$

leading to the solution

$$\widehat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Note that here we have assumed that $\mathbf{X}^T\mathbf{X}$ is a non-singular matrix. If the design matrix $\mathbf{X}^T\mathbf{X}$ is singular, we can use a generalized inverse of $\mathbf{X}^T\mathbf{X}$.

The Gauss-Markov theorem states that OLS estimator has the minimum variance among any other unbiased linear estimators. However, OLS method suffers in prediction particularly if $\mathbf{X}^T\mathbf{X}$ is singular or near singular. The prediction error can be improved by reducing variance but allowing some bias. Ridge regression and LASSO regression are penalization techniques applied to the model.

Ridge regression, introduced by Hoerl and Kennard (1970), penalizes the size of the regression coefficient. The estimate is obtained by minimizing the penalized sum of squares,

$$\text{SSE } (\boldsymbol{\beta}, \lambda) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}, \tag{2.5}$$

where λ is a complexity parameter that controls the amount of shrinkage. If λ increases to infinity, the amount of shrinkage is larger. Now the solution of (2.5) is,

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{2.6}$$

where $\mathbf{I}$ is the $(p+1) \times (p+1)$ identity matrix. Equivalently the ridge coefficient minimizes the usual SSE in (2.4) subject to the constraint that $\sum_{j=0}^{p} \beta_j^2 \leq s$, where s is the parameter playing the same role of λ. The difference between the ridge coefficients and OLS coefficients is λ times the identity matrix added to $\mathbf{X}^T\mathbf{X}$. In fact, $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})$ is not singular even if $\mathbf{X}^T\mathbf{X}$ is a singular matrix. For $\mathbf{X}^T\mathbf{X}$ nearly singular, $\text{Var}(\widehat{\boldsymbol{\beta}}_{\text{ridge}})$ can be substantially lower than $\text{Var}(\widehat{\boldsymbol{\beta}}_{\text{OLS}})$. This is the main motivation behind ridge regression.

Ridge coefficients minimize SSE subject to a bound L_2-norm of the coefficients.

LASSO, introduced by Tibshirani (1996), minimizes SSE subject to a bound on the L_1-norm. Unlike the ridge regression, the LASSO can make some coefficients exactly zero so that we can choose a subset of the predictors which have relatively stronger effect on the dependent variable Y. There is a discrete variable selection procedure so called best subset. However, the LASSO method works both as a continuous shrinkage and an automatic variable selection technique. Similar to the ridge method, the LASSO coefficients can be defined that minimizes the SSE subject to $\sum_{j=0}^{p} |\beta_j| \leq s$.

Tibshirani (1996) and Fu (1998) compared above methods in terms of prediction error but there is no one which is superior to the other method. Recently, Zou and Hastie (2005) developed a method which combined the ridge regression and LASSO to give what is called the "elastic net (EN)" and used in microarray data analysis. Especially EN outperforms the LASSO when the number of predictors, p, is much larger than the number of observations, n, ("the large p, small n" problem) and when there is a group of highly correlated variables.

In our problem, if there are unknown clusters we want to apply above regression methodologies at the second step, given several candidate clusters. In low dimension ($p \leq 4$, say), variable selection may not be important. For higher dimension ($p > 4$), variable selection is necessary to avoid overfitting and to achieve stability. In such cases, we propose using the LASSO regression, which automatically selects a simpler model.

2.1.2 Nonparametric regression

The linear regression model has the advantages of simple interpretation and easy computation. But, it has a limited flexibility and produces accurate estimates only when the true form of the regression function is close to the linear function. This motivates the use of more flexible models.

Univariate case

On estimation of nonparametric regression function there are roughly three paradigms such as local parametric, piecewise fitting and roughness penalty method.

- Local parametric fitting : There are two main issues in this method. One is how big the neighborhoods are and the other is how to average the target value in each neighborhood. Bin smoother, which is also known as "regressogram" is a basic method like histogram for density estimation. If we use the least square method in a neighborhood then the running line method is obtained. A useful and popular estimator is the kernel estimator proposed by Nadaraya (1964) defined by

$$\widehat{f}(X) = \frac{\sum_{i=1}^{n} K_h(X - X_i) Y_i}{\sum_{k=1}^{n} K_h(X - X_k)}, \tag{2.7}$$

where $K_h(\cdot) = \frac{1}{h} K(\frac{\cdot}{h})$ is a kernel function having bandwidth $h > 0$ and $K(\cdot)$ is a symmetric density function. It is useful to represent this in matrix notation.

$$\widehat{f}(\mathbf{X}) = \mathbf{S}\mathbf{Y}, \quad \text{where } S_{i,j} = \frac{\sum_{i=1}^{n} K_h(X_j - X_i)}{\sum_{k=1}^{n} K_h(X - X_k)}, \tag{2.8}$$

and $\mathbf{X}$ and $\mathbf{Y}$ are vector of length n. Then, the $n \times n$ matrix $\mathbf{S}$ is called smoother

matrix. For the method of least squares, the smoother matrix is nothing but the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. This representation also applies to smoothing splines and other nonparametric methods.

There may be many choices for the kernel functions but it is known that the choice of bandwidth h is more important than that of the kernel function. A common choice for the kernel function is the standard normal distribution. A common method for bandwidth selection is that of cross validation (CV). For given data $(X_1, Y_1), \ldots, (X_n, Y_n)$, for $i = 1, \ldots, n$, we leave (X_i, Y_i) out and fit the estimate of f, say $\widehat{f}_{(-i)}$ then calculate the error as

$$CV = \frac{1}{n}\sum_{i=1}^{n}\Big(Y_i - \widehat{f}_{(-i)}(X_i)\Big)^2 = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{Y_i - \widehat{f}(X_i)}{1 - S_{i,i}}\Big)^2,$$

where $S_{i,i}$ is the diagonal elements of the smoother matrix $\mathbf{S}$. Note here, $\widehat{f}_{(-i)}$ and $\widehat{f}$ depend on the bandwidth h. The optimal h can be obtained by minimizing CV. Alternatively generalized cross validation (GCV) is proposed by replacing $S_{i,i}$ by $\frac{1}{n}\,\mathrm{tr}(S)$. Craven and Wahba (1979) proposed GCV method and established the optimal property of GCV. In this thesis, we will use the method which was proposed by Sheather and Jones (1991). This is already implemented in R package, "KernSmooth" with function "dpik" and "ksmooth".

- Piecewise fitting : This method basically divides the domain region into several intervals and approximate the regression function with low order of polynomials within each interval. The cut-off points necessary for this method are called knots. Continuity at the all the knots is imposed. The usual method is the polynomial regression spline method. The cubic spline approximates the third

order of polynomial at each subregion. The main difficulty is to select the order of polynomial as well as the knots. Splines can be represented by different basis functions. For example, cubic regression splines can be expressed by the following.

$$\widehat{f}(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=1}^{N} \theta_j (x - \xi_j)_+^3, \tag{2.9}$$

where $\xi_1, \ldots, \xi_N$ are given knots and $(a)_+ = \max(a, 0)$.

- Roughness penalty method : A Sobolev space is defined as a normed space of functions for which derivatives upto a certain order satisfy some integrability restrictions. Let $f(x)$ be a function in the second order Sobolev space on unit interval, $W_2[0, 1]$, then a smoothing spline is a function $f \in W_2[0, 1]$ which minimizes

$$\frac{1}{n} \sum_{i=1}^{n} \Big(Y_i - f(X_i) \Big)^2 + \lambda \int_0^1 \Big(f''(x) \Big)^2 dx, \tag{2.10}$$

where $\lambda > 0$ is a smoothing parameter. The second term of (2.10) penalize the roughness or the curvature of estimator. The parameter λ controls the bias-variance trade off. For example, if $\lambda = 0$, the $\widehat{f}(X_i) = Y_i$ and if $\lambda = \infty$ then $\widehat{f}$ is the least square regression line. The solution of (2.10) is called cubic spline with knots $X_1, \ldots, X_n$. It is important to choose λ properly. The function "smooth.spline" in R, uses CV method to find such λ and fit the resulting function.

Multivariate case

Nonparametric regression method is challenged in the multidimensional case because of the curse of dimensionality, which means that in high dimensions, a neighborhood containing a fixed percentage of data points can be very big. There have been many studies to overcome this difficulty. Mostly, a dimension reduction technique is considered and then smoothing methods are used even though there are direct smoothing approaches such as thin-plate spline and classification and regression trees (CART). Here we review some of recent methods such as the generalized additive model (GAM) and multivariate adaptive regression splines (MARS).

The additive model is a generalization of the usual linear regression model, where the regression function is assumed to be a sum of functions on lower-dimensions. The additive model is defined as

$$Y_i = \sum_{j=1}^{p} f_j(X_{j,i}) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.11}$$

where $\varepsilon_i \overset{iid}{\sim} N(0, T^2)$. Note that f_j's are arbitrary univariate functions, one for each predictor. Hastie and Tibshirani (1990) adapted this additive model to generalized linear models, and this is called the generalized additive model (GAM). The GAM differs from a generalized linear model in that an additive part replaces the linear predictor. The GAM is defined as

$$g(\mu) = \sum_{j=1}^{p} f_j(X_j), \tag{2.12}$$

where the mean $\mu(X) = E(Y|X)$ is linked through $g(\cdot)$. The estimation method in the additive model is known as the "back-fitting" method. For the GAM, the

local scoring procedure plays a role as the iteratively reweighted least square method in a generalized linear model. In R, "gam" function is implemented in the package "mgcv". Since it is necessary to use the nonparametric estimation method to estimate each function f_j, the bandwidth parameters are selected by some suitable data driven methods. In R, "gam" function uses the generalized cross validation method to select the bandwidth parameter.

MARS is a method for flexible modeling in high dimensional data, motivated from the recursive partitioning model by Friedman (1991) . This does not assume any parametric form of $f(X)$. MARS constructs the function from a set of coefficients and basis functions only from the data. The model is defined by

$$Y_i = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(X_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.13}$$

where B_m is a basis function. For example, suppose we consider the class,

$$\mathfrak{C} = \{(X_j - t)_+, (t - X_j)_+ | t \in \{X_{1,j}, \ldots, X_{n,j}\}, j = 1, \ldots, p\},$$

where $(s)_+ = s$ if $s > 0$, 0 otherwise. Then B_m is either a function in $\mathfrak{C}$ or a product of two or more such functions. By allowing the order of interactions, the above model can be an interactive model instead of the additive one. In R, this can be controlled by the option "degree" in the function "mars".

2.2 Bayesian Methods

2.2.1 Gaussian process method

A Gaussian process is a collection of random variables and the finite number of which have joint normal distributions. A Gaussian process can be specified by the mean function $m(x)$ and covariance function $K(x, x')$. Let $f(x)$ denote the regression function distributed as a Gaussian process (GP) with mean function $m(x)$ and covariance function $K(x, x')$, then $f \sim GP(m, K)$. The choice of an appropriate covariance function enables to have a large support in the space of all smooth functions. For example, the covariance function $K(x, x') = \frac{1}{\tau} e^{-\gamma(x-x')^2}$ with varying γ spans the space of all smooth functions. The weak but crucial assumption for the covariance function is that it generates non-negative definite matrix for any set of inputs.

Suppose we have given data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and we want to predict the estimate of $f(x_{n+1})$ at given x_{n+1}. Let $\mathbf{f} = (f(X_1), \ldots, f(X_n))^T$ and $\mathbf{m} = (m(X_1), \ldots, m(X_n))^T$ and Σ_0 the covariance matrix having the $(i, j)^{th}$ element $K(X_i, X_j)$. The conditional distribution of f given data is

$$f|(X_1, Y_1), \ldots, (X_n, Y_n) \sim GP(m^*, K^*), \tag{2.14}$$

where

$$m^*(x) = m(x) + K(x, \mathbf{X})^T \Sigma_0^{-1} (\mathbf{f} - \mathbf{m}),$$

$$K^*(x, \mathbf{X}) = K(x, x) - K(x, \mathbf{X})^T \Sigma_0^{-1} K(x, \mathbf{X}),$$

and $K(x, \mathbf{X}) = (K(x, X_1), \ldots, K(x, X_n))^T$. The above expression is easily obtained by the fact that random vectors $f(X_1), \ldots, f(X_n), f(x_{n+1})$ are distributed as multi-

variate normal distribution, and the distribution of $f(x_{n+1})$ is the conditional distribution from the joint normal.

Now the specification of hyper parameters such as putting a prior for τ, γ in above example covariance function gives us posterior updating formula. An MCMC method is considered by Neal (1997) and MacKay (1999). One difficulty for this method is its long computing time, as the covariance function, which depends on the data size needs to be inverted. Before modern cheap and fast computing machine it was not feasible to compute the inverse of matrix whose dimension is more than hundred. But, now it is feasible to apply this method in the case of more than thousands of samples. The strength of GP method is the conceptual simplicity and the flexibility in modeling. Rasmussen (1996) has compared GP with other nonparametric regression methods and shown that GP has better performance among them. Geostatistics is a common application area of GP, where the predictor is of two or three dimensions [Cressie (1993)]. Choi and Schervish (2004) established the posterior consistency of Gaussian process priors for the regression function in normal error case in one dimension.

2.2.2 Basis expansion method

Many Bayesian methods for nonparametric regression problem use some basis expansion such as splines, the Fourier basis and wavelet bases. Suppose that given a basis $\{f_1, \ldots, f_M\}$, the regression function can be expressed by

$$f(\cdot) = \sum_{j=1}^{M} b_j f_j(\cdot), \tag{2.15}$$

where $b = (b_1, \ldots, b_M)$ is a vector of basis coefficients.

Splines commonly use basis functions as

$$\{1, x, x^2, x^3, (x-\xi_1)_+^3, \ldots, (x-\xi_p)_+^3\}, \tag{2.16}$$

where $(x)_+ = \max(x, 0)$ and $\xi = (\xi_1, \ldots, \xi_p)$ is a vector of knots. For a given ξ, the regression function can be formed by the coefficients $b = (b_1, \ldots, b_M)$. The specification of prior for (ξ, b, σ) enable to set up a Bayesian model in this problem. There are many approaches depending on different priors on b (see Section 3.1 in Muller and Quintana (2004)). Updating ξ, a vector of knots, is somewhat difficult when the number of knots and locations are unknown. Denison *et al.* (1998) proposed a procedure which enables to add, delete and change knots from the data by using the reversible jump MCMC method. For the multivariate case, the additivity assumption is common. Denison *et al.* (1998) assumed the additive structure. Alternatively, Denison *et al.* (1998b) proposed Bayesian approach to MARS by using the reversible jump MCMC method. They showed that Bayesian MARS have a high predictive power due to posterior averaging.

Wavelet basis methods provide an orthonormal L_2 basis of the space of square integrable functions. Any L_2 function f can be represented by the wavelet expansion,

$$f(x) = \sum_j \sum_k d_{j,k} \psi_{j,k}(x), \tag{2.17}$$

with basis function $\psi_{j,k}(x) = 2^{\frac{j}{2}} \psi(2^j x - k)$. This representation is usually very sparse so that there are only a few coefficients having relatively large values. The idea is to first consider the discrete wavelet transform from given data and then discard those small coefficients by thresholding. With those coefficients, one can reconstruct the original data by the inverse of discrete wavelet transformation. Let $d_{j,k}^*$ denote the

empirical coefficients obtained by applying discrete wavelet transform to the data $(Y_1, \ldots, Y_n)$. Then, Bayesian wavelet regression methods need to get the posterior distribution of $d_{j,k}|d^*_{j,k}$ by putting a prior on $d_{j,k}$. Chipman *et al.* (1997) put an independent prior on $d_{j,k}$,

$$d_{j,k}|\gamma_{j,k} \sim \gamma_{j,k}N(0, c_j^2\nu_j^2) + (1-\gamma_{j,k})N(0, \nu_j^2),$$

where $\gamma_{j,k}$ ∼Bernoulli(p_j) and all the hyperparameters are determined by empirical methods. Then, the posterior distribution is also a mixture of two normal distributions because d^* has a normal distribution with mean $d_{j,k}$ and a fixed variance.

Abramovich *et al.* (1998) placed independent priors on $d_{j,k}$ as

$$d_{j,k} \sim p_j N(0, \tau_j^2) + (1-p_j)\delta(0),$$

where $\delta(0)$ is degenerate probability at 0. Vidakovic (1998) proposed a mixture of a point mass and a $t-$distribution. Holmes and Denison (1999) considered an infinite mixture of normals.

2.3 Other methods

DeSarbo and Cron (1988) proposed a conditional mixture maximum likelihood methodology for performing clusterwise linear regression. The term of "clusterwise regression" was first used by Spath (1979). First, it needs to use AIC (Akaike's information criterion) in order to estimate the number of cluster k. Then, given k, the weighted linear regression function can be estimated by using EM algorithm.

Muller, Erkanli and West (1996) suggested a nonparametric Bayesian curve fitting method. Their idea is that once the distribution of joint density $(\boldsymbol{X}, Y)$ is known then the conditional distribution of $Y|\boldsymbol{X}$ can be easily obtained. The joint distribution of $(\boldsymbol{X}, Y)$ is assumed to be a mixture. Dirichlet mixtures of normals prior was used to estimate the joint distribution. Their method is a fully Bayesian method. The method suffers in higher dimensional case and the large p small n problem and also when the conditional mean is not linear in each subpopulation.

Geweke and Keane (2005) developed a flexible model called smoothly mixing regression method. They use the multinomial probit model for the posterior probability of unknown cluster given covariates. The main reason of using this model is that it produces simple and tractable posterior distributions in Gibbs sampling algorithm. They give two illustrations for their method, earning data and stock return data. There are two important difficulties. One is that the evaluating the conditional probability density is somewhat awkward. The other is a well known difficulty in high dimension, the curse of dimensionality in the order of the conditioning covariates.

Chapter 3

Preliminaries

In this chapter we review Bayesian nonparametric method concepts such as the Dirichlet process, and the Dirichlet mixture process and their properties. The posterior is typically computed using the MCMC method of Gibbs sampling. The MCMC sampling scheme and the consistency of posterior distribution will be reviewed in this chapter.

3.1 Dirichlet process and Dirichlet mixture process

Nonparametric inference is concerned about an unknown parameter of infinite dimension. Functions such as density distribution, regression function or survival functions may be of interest in an inference problem. Bayesian methods usually require prior information about the unknown parameters. Therefore, Bayesian nonparametrics requires the construction of a prior on infinite dimensional space such as the space of probability measure. A popular prior, which is a stochastic process

whose sample paths are probability measures, is the Dirichlet process introduced by Ferguson (1973).

Let $\mathfrak{M}(\mathfrak{X})$ be the space of probability measure on $\mathfrak{X}$. Let G be a probability measure on $(\mathfrak{X}, \mathcal{B})$ where $\mathcal{B}$ is the Borel σ-field.

Definition 1 *Let $M > 0$ and G be a probability measure on $(\mathfrak{X}, \mathcal{B})$. A random probability measure P follows a Dirichlet process on $(\mathfrak{X}, \mathcal{B})$ if for any finite partition $\{B_1, \ldots, B_k\}$ of $\mathfrak{X}$, the joint distribution of the vector $(P(B_1), \ldots, P(B_k))$ has the Dirichlet distribution with parameter $(MG(B_1), \ldots, MG(B_k))$.*

The Dirichlet distribution for variables $(x_1, \ldots, x_k)$ with parameter $(p_1, \ldots, p_k)$ is defined by

$$f(x_1, \ldots, x_k) = \frac{1}{C} \prod_{i=1}^{k} x_i^{p_i - 1},$$

when $x_1, \ldots, x_k \geq 0$, $\sum_{i=1}^{k} x_i = 1$ and $p_1, \ldots, p_k > 0$. The constant C is given by $\prod_{i=1}^{k} \Gamma(p_i) / \Gamma(\sum_{i=1}^{k} p_i)$, where $\Gamma(\cdot)$ is the gamma function. We denote "$DP(M, G)$" as the Dirichlet process with parameter (M, G). Note here,

$$\mathrm{E}(P(B)) = G(B), \quad \mathrm{Var}(P(B)) = \frac{G(B)(1 - G(B))}{1 + M}. \tag{3.1}$$

If M gets larger, then P approaches G. The number M is called the precision parameter and G is called the center measure, and MG is referred to as the base measure of $DP(M, G)$. An attractive mathematical property of Dirichlet process is that given a set of realizations $\theta_1, \ldots, \theta_n$ from $P \sim DP(M, G_0)$, the posterior distribution of P is also a $DP(M^*, G_0^*)$ where $M^* = M + n$ and $G_0^* = (MG_0 + \sum_{i=1}^{n} \delta_{\theta_i})/(M + n)$; here $\delta_\theta(\cdot)$ is the indicator function, $I\{x = \theta\}$. This can be illustrated by the Polya urn

scheme model which also gives a construction of the Dirichlet process [Blackwell and MacQueen (1973)].

$$\begin{aligned} \theta_1 &\sim G_0 \\ \theta_i|\theta_1,\ldots,\theta_{i-1} &\sim \frac{MG_0 + \sum_{j=1}^{i-1}\delta_{\theta_j}}{M+i-1}, \quad \text{for } i = 1,2,\ldots. \end{aligned} \tag{3.2}$$

Therefore the predictive distribution of θ_{n+1} can be derived by the following

$$\theta_{n+1}|\theta_1,\ldots,\theta_n \sim \frac{M}{M+n}G_0 + \frac{1}{M+n}\sum_{i=1}^{n}\delta_{\theta_i}. \tag{3.3}$$

Sethuraman (1994) gave a useful construction of Dirichlet process. Suppose P is from $DP(M,G)$. Then

$$P = \sum_{i=1}^{\infty} V_i \delta_{\theta_i}, \tag{3.4}$$

where $\theta_1, \theta_2, \ldots$ are a sequence i.i.d. random variable from G, $V_i = Y_i \prod_{j=1}^{i-1}(1-Y_j)$ and $Y_1, Y_2, \ldots$ are i.i.d. random variables from $Beta(1,M)$. Thus probabilities are assigned by "stick breaking" at randomly distributed points. It can be easily shown that $\sum_{i=1}^{\infty} V_i = 1$ almost surely. With this definition of Dirichlet process, we can generate a realization of the Dirichlet process by truncation at some finite stage. It follows that Dirichlet process is almost surely discrete.

The weak support of Dirichlet process, $DP(M,G)$ can be shown to be $\{P \in \mathfrak{M}(\mathfrak{X}) : \text{supp}(P) \subset \text{supp}(G)\}$. This means that if the support of G is $\mathfrak{X}$, then the space of all probability measures is the support of P. For example, if we have normal distribution as G, then the Dirichlet process can choose any probability measure.

Since the Dirichlet process puts all its mass on the subset of all discrete distributions, smoothing the Dirichlet process is particularly important for density estimation.

Suppose that $X_1, \ldots, X_n$ are drawn as a random sample from the distribution of

$$f(\cdot, G) = \int K(\cdot, \theta, \sigma) dG(\theta), \tag{3.5}$$

where $K(\cdot, \theta, \sigma)$ is a density function given θ, σ and $G \sim DP(M, G_0)$. This was developed by Ferguson (1983) and Lo (1984), and is called "Dirichlet mixture process". The kernel $K(\cdot, \theta, \sigma)$ can be any density function. The normal distribution with mean θ and variance σ^2 is a common choice and we will work with this one. Note that we can put the Dirichlet prior on the distribution of θ or (θ, σ^2). In the latter case, the center distribution G_0 is the joint distribution of (θ, σ^2). If we consider the Dirichlet prior on the mean θ, then the normal Dirichlet mixture process model is

$$f(\cdot, G) = \int N(\cdot; \theta, \sigma^2) dG(\theta), \tag{3.6}$$

where $N(\cdot; \theta, \sigma^2)$ is the p.d.f. of normal distribution with mean θ and variance σ^2. There is an alternative useful representation in terms of a hierarchical model, given below:

$$\begin{aligned} \boldsymbol{X_i} | \theta_i &\overset{ind}{\sim} N(\cdot; \theta_i, \sigma^2), \\ \theta_i | G &\overset{iid}{\sim} G, \\ G &\sim DP(M, G_0). \end{aligned} \tag{3.7}$$

In Bayesian inference, we get the posterior distribution of all the parameters that we are interested in. The joint distribution of $(\theta_1, \ldots, \theta_n, \sigma^2)$ is not analytically tractable, so the Gibbs sampling technique is considered for which we need to sample iteratively from one dimensional conditional distributions. More precisely, the idea of Gibbs sampling scheme can be summarized as follows:

Suppose we are interested in sampling $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ from the distribution $\pi(\boldsymbol{\theta})$. Let $\theta_{-i} = \{\theta_j; j \neq i\}$ for $i = 1, \ldots, n$. Assume that the full conditional distributions $\pi(\theta_i|\theta_{-i})$ are given and it is easy to sample from it.

(i) Set initial values $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_n^{(0)})$.

(ii) Obtain a new value $\boldsymbol{\theta}^{(j)} = (\theta_1^{(j)}, \ldots, \theta_n^{(j)})$ from $\boldsymbol{\theta}^{(j-1)}$ through succesive sampling

$$\begin{aligned} \theta_1^{(j)} &\sim \pi(\theta_1|\theta_2^{(j-1)}, \ldots, \theta_n^{(j-1)}), \\ \theta_2^{(j)} &\sim \pi(\theta_2|\theta_1^{(j)}, \theta_3^{(j-1)} \ldots, \theta_n^{(j-1)}), \\ &\vdots \\ \theta_n^{(j)} &\sim \pi(\theta_n|\theta_1^{(j)}, \ldots, \theta_{n-1}^{(j)}). \end{aligned} \tag{3.8}$$

(iii) Increase the counter j to $j+1$ and repeat the step (ii) until certain convergence rule is satisfied.

The Gibbs sampler is a very powerful algorithm and its implementation is often easy in many complex problems. According to the Gibbs sampler, we only need to obtain the conditional distribution of θ_i given $\theta_{-i}, \boldsymbol{X}$ and σ^2. The following gives the basic idea on how to implement MCMC steps in a Dirichlet mixture model.

Algorithm (1).

$$\theta_i|\theta_{-i}, \sigma^2, \boldsymbol{X} \propto q_{i0} G_i(\theta_i) + \sum_{j \neq i} q_{ij} \delta_{\theta_j}(\theta_i), \tag{3.9}$$

where

$$q_{i0} \propto \int N(\boldsymbol{X_i}, \theta, \sigma^2) dG_0(\theta),$$

$$q_{ij} \propto N(\boldsymbol{X_i}; \theta_j, \sigma^2),$$
$$\sum_{j=1}^{n} q_{ij} + q_{i0} = 1,$$
$$G_i(\theta_i) \propto N(\boldsymbol{X_i}; \theta, \sigma^2) dG_0(\theta).$$

Since there are many ties during the updating of θ_i, the above algorithm is not efficient and an alternative algorithm is usually considered. Let ϕ be the set of distinct θ_i's and let k be the number of distinct elements of $\theta_1, \ldots, \theta_n$. Let $s = (s_1, \ldots, s_n)$ be the configuration vector defined by

$$s_i = j \text{ if and only if } \theta_i = \phi_j, \quad i = 1, \ldots, n \quad j = 1, \ldots, k.$$

Let $H = \{I_1, \ldots, I_k\}$ be a cluster which is defined by

$$I_j = \{i : s_i = j\}.$$

Therefore, H is a partition of $I = \{1, \ldots, n\}$. Now given the configuration vector s and ϕ, θ and H are uniquely determined by the following rule.

$$\begin{aligned} \phi_1 &= \theta_1 \\ \phi_j &= \theta_i, \text{ if } j \geq 2 \text{ and } i = \min\{m : \theta_m \neq \phi_1, \ldots, \theta_m \neq \phi_{j-1}\} \end{aligned} \tag{3.10}$$

Let the following be the notations when the observation i is removed:

- $\theta_{-i} = \{\theta_j : 1 \leq j \leq n, j \neq i\}$,
- k_{-i}: the number of clusters formed by θ_{-i},
- ϕ_{-i} : the set of distinct observations among θ_{-i},
- $n_{-i,j}$: the number of elements in cluster j when the i^{th} observation is removed.

The following is the alternative simplified algorithm for (3.9);

$$(\theta_i|\theta_{-i},\sigma^2,\boldsymbol{X}) \propto q_{i0}G_i(\theta_i) + \sum_{\phi_k\in\phi_{-i}} n_{-i,k}q_{ik}\delta_{\phi_k}(\theta_i), \tag{3.11}$$

where $q_{ik} \propto N(\boldsymbol{X_i}|\phi_k,\sigma^2)$.

It is known that above algorithm is not efficient because there is a small chance to have a new value of θ during procedure. There is a technique of "remixing" the ϕ_j after every step to prevent this problem. The vector of $(\theta_1,\ldots,\theta_n)$ can be fully determined by knowing the configuration vector $(s_1,\ldots,s_n)$ and distinct values $(\phi_1,\ldots,\phi_k)$. Therefore sampling θ_i is equivalent to sampling s_i and ϕ_i given s,ϕ_{-i}, $i=1,\ldots,k$. Note here that if $s_i = j \le k_{-i}$, then the new $\theta_i = \phi_j$ and if $s_i = k_{-i}+1$, then let the new value θ_i be sampled from G_i. This gives an alternative and more efficient updating procedure.

Algorithm (2).

Under the notation of (3.11), the distribution of s_i is

$$\begin{aligned} P(s_i = j|s_{-i},\phi_{-i},\boldsymbol{X}) &\propto n_{-i,j}N(\boldsymbol{X_i};\phi_j,\sigma^2), \\ P(s_i = k_{-i}+1|s_{-i},\phi_{-i},\boldsymbol{X}) &\propto M\int N(\boldsymbol{X_i};\theta,\sigma^2)dG_0(\theta), \end{aligned} \tag{3.12}$$

and the posterior distribution of $\phi_1,\ldots,\phi_k$ are

$$P(\phi_j|s,\boldsymbol{X}) \propto dG_0(\phi_j)\prod_{i\in I_j} N(\boldsymbol{X_i};\phi_j,\sigma^2) \quad \text{for } j=1,\ldots,k. \tag{3.13}$$

3.2 Evaluation of clustering

In the previous section, our first step is to find the clustering among data by using the Dirichlet mixture model. The key idea is that the clustering obtained

from Dirichlet process can give us a "good" partition. Now the question is how the clustering is good? According to Rand (1971), an objective criterion should have the following three basic properties. First, clustering is discrete, that is every point is assigned to a specific cluster. Second, clusters are defined just as much by those points which they do not contain as by those points which they do contain. Third, all points are of equal importance in the determination of clustering.

Given n points, $\theta_1, \ldots, \theta_n$ and two clustering vectors, $\mathbf{s} = \{s_1, \ldots, s_{k_1}\}$ and $\mathbf{s}' = \{s'_1, \ldots, s'_{k_2}\}$, define a measure d be defined by

$$d(\mathbf{s}, \mathbf{s}') = \frac{\sum_{i<j}^{n} \gamma_{ij}}{\binom{n}{2}}, \tag{3.14}$$

where

$$\gamma_{ij} = \begin{cases} 1, & \text{if there exist } k \text{ and } k' \text{ such that both } \theta_i \text{ and } \theta_j \text{ are in both } s_k \text{ and } s'_{k'}, \\ 1, & \text{if there exist } k \text{ and } k' \text{ such that } \theta_i \text{ is in both } s_k \text{ and } s'_{k'} \text{ while } \theta_j \\ & \text{is in neither } s_k \text{ or } s'_{k'}, \\ 0, & \text{otherwise.} \end{cases} \tag{3.15}$$

In practice, we use a simple computational formula for d:

$$d(\mathbf{s}, \mathbf{s}') = \Big[\binom{n}{2} - 0.5\Big(\sum_i (\sum_j n_{ij})^2 + \sum_j (\sum_i n_{ij})^2\Big) + \sum\sum n_{ij}^2\Big] \Big/ \binom{n}{2}, \tag{3.16}$$

where n_{ij} is the number of points simultaneously in the i^{th}cluster of $\mathbf{s}$ and the j^{th} cluster of $\mathbf{s}'$. This is the measure of similarity and ranging from 0 to 1. When $d = 0$ the two clusterings have no similarities, and when $d = 1$ the clusterings are identical. In this thesis, this measurement of similarity is used in MCMC steps. We call it "Rand's measure".

3.3 Inference about M

In MCMC updating scheme, a suitable choice of M is important because M controls the number of clusters, k. We may put a value for M as an initial moderate choice and a prior on M lets the data choose the appropriate M. In this section we review the distribution of k. The aim is to give the posterior distribution of M so that we can update during MCMC procedure. This approach originally came from Escobar and West (1995). There are alternative ways to put a prior on M. Escobar (1994, 1998) used a discrete probability on a suggested grid points. Liu (1996) showed how to obtain the maximum likelihood estimate of M.

The distribution of the number of component k is the following:

$$P(k|M,n) = C_n(k) \cdot n! \cdot M^k \cdot \frac{\Gamma(M)}{\Gamma(M+n)}, \tag{3.17}$$

where $C_n(k) = P(k|M=1,n)$. The expectation of k is given by

$$\mathrm{E}(k) = \sum_{i=1}^{n} \frac{M}{M+i-1} \approx M \log\Big(\frac{M+n}{M}\Big). \tag{3.18}$$

Then, we can obtain the posterior of M given data $\boldsymbol{X}$,

$$\begin{aligned}
P(M|k,\theta,\boldsymbol{X}) &= P(M|k) \quad (\because \text{the independence of } M, \theta, \text{ and } \boldsymbol{X} \text{ given } k) \\
&\propto P(M) \cdot P(k|M) \\
&\propto P(M) \cdot n! \cdot M^k \cdot \frac{\Gamma(M)}{\Gamma(M+n)} \quad (\because C_n(k) \text{ does not depend on } M) \\
&\propto P(M) M^k B(M,n) \\
&= P(M) M^k \int_0^1 \eta^{M-1}(1-\eta)^{n-1} du,
\end{aligned} \tag{3.19}$$

where $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$. Then, $P(M|k)$ can be considered as the marginal distribution from a joint for M and a continuous quantity η such that

$$P(M,\eta|k) \propto P(M) \cdot M^k \eta^{M-1}(1-\eta)^{n-1}, \quad M > 0, \quad 0 < \eta < 1.$$

Thus, if $P(M) = Gamma(a,b)$, then the posterior of M is

$$\begin{aligned} P(M|\eta,k) &\propto M^{a-1}e^{-bM}M^k\eta^{M-1} \\ &\propto M^{a+k-1}e^{-M(b-\log\eta)} \\ &= Gamma(a+k, b-\log\eta). \end{aligned} \tag{3.20}$$

These distributions are well defined for all gamma priors and

$$P(\eta|M,k) \propto \eta^{M-1}(1-\eta)^{n-1}, \quad 0 < \eta < 1.$$

Therefore in MCMC sampling steps, repeating to sample η and M will give us an appropriate value of M. Now, the estimated posterior distribution of M can be obtained by the following:

$$P(M|\text{Data}) \approx \frac{1}{N}\sum_{i=1}^{N} P(M|\eta_i, k_i) \tag{3.21}$$

where η_i are the sample values of η.

3.4 Posterior consistency

In this section we briefly review the literature on the consistency of the posterior distribution. In Bayesian inference, consistency is important because a posterior may be misled towards a wrong value without it. The posterior is consistent if it concentrates near the true value as more data are collected. This means that the information

from the data eventually dominates whatever prior information was. We will briefly illustrate some preliminary definitions and give an argument which indicates posterior consistency in the Dirichlet mixture model. Formally speaking, let Π_n be the posterior given samples $X_1, \ldots, X_n$. Then Π_n is consistent at θ_0 if $\Pi_n(U) \to 1$ a.s. for every neighborhood U of θ_0. It is equivalent to $\Pi_n \xrightarrow{w} \delta_{\theta_0}$, "$w$" denotes convergence of measures in the weak sense. Naturally, posterior consistency depends on the topology on the parameter space.

A fundamental theorem is given by Schwartz (1965) but it seems to be difficult to apply her theorem to our regression problem directly. We shall extend the consistency property in Dirichlet mixture of normal model of Ghosal et al. (1999) to the multivariate case. At present the weak consistency result is only proved. A result from Donoho (1988) showed that the number of mixture is a lower semi-continuous functional indicating that asymptotically the number of clusters produced by the Dirichlet mixture is at least equal to the number of components in the true mixture distribution. This will be discussed in Chapter 4.

Consistency under two topologies, the weak topology and the strong topology will be discussed in this chapter. We first give some definitions of neighborhoods on the space of all density functions. Let $\mathfrak{F}$ be the space of all densities on $\mathbb{R}$ with respect to the Lebesgue measure. On $\mathfrak{F}$, the weak topology and the norm topology can be considered as natural topologies. Since a topology can be formed by defining the system of neighborhoods, we can start by giving two natural types of neighborhoods.

Definition 2 *If $f_0 \in \mathfrak{F}$, a weak neighborhood of f_0 is a set containing a set of the*

form:

$$\Big\{f \in \mathfrak{F} : \Big| \int \psi_i f - \int \psi_i f_0 \Big| < \varepsilon \text{ for } i = 1, \ldots, k\Big\}, \tag{3.22}$$

where ψ_i are bounded continuous functions on $\mathbb{R}$. This topology is called "weak topology".

A strong neighborhood is a set containing a set of the form:

$$\{f \in \mathfrak{F} : \|f - f_0\| < \varepsilon\},$$

where $\|\cdot\|$ denotes a metric.

Definition 3 *A prior Π is said to be weakly consistent at f_0, if*

$$\Pi(U|X_1, \ldots, X_n) \to 1$$

with probability 1 for all weak neighborhoods U of f_0.

The closeness between two functions can be measured by a metric, norm and so on. One such measure is the Kullback-Leibler divergence.

Definition 4 (K-L divergence) *The Kullback-Leibler divergence between two densities f, f_0 is given by*

$$K(f_0, f) = \int f_0(x) \log \frac{f_0(x)}{f(x)} dx.$$

Definition 5 (K-L support) *For any $f_0 \in \mathfrak{F}$, the Kullback-Leibler neighborhood of radius ε around f_0, is given by $K_\varepsilon(f_0) = \{f : K(f_0, f) < \varepsilon\}$.*

Let Π be a prior on $\mathfrak{F}$, we say that f_0 is in the K-L support of Π if

$$\Pi(K_\varepsilon(f_0)) > 0 \quad \text{for all } \varepsilon > 0.$$

Theorem 1 [Schwartz (1965)]. *If f_0 is in the K-L support of Π, then the posterior is weakly consistent at f_0.*

For the strong consistency, Barron, Schervish and Wasserman (1999) used the Hellinger bracketing entropy while Ghosal *et al.* (1999) used L_1-metric entropy which gives a condition weaker than Barron *et al.* (1999).

Theorem 2 [Ghosal et al. (1999)]. *Suppose f_0 is in the K-L support of the prior Π and for every $\epsilon > 0$ there exists a $\delta < \epsilon, c_1, c_2 > 0, \beta < \epsilon^2/2$ and $\mathcal{F}_n$ such that*

(i) $\Pi(\mathcal{F}_n^c) < c_1 e^{-nc_2}$,

(ii) $J(\delta, \mathcal{F}_n) < n\beta$, *where $J(\delta, \mathcal{F}_n)$ is the logarithm of the minimal number of balls of radius δ in the total variation metric needed to cover the $\mathcal{F}_n$.*

Then the posterior is consistent in the total variation distance.

Because our regression method is based on the estimation of the underlying mixing distribution which is an inverse problem, establishing consistency seems to be difficult at present. In this thesis, we have not attempted to prove consistency explicitly, but we do give some useful asymptotic results which indicate the plausibility of consistency of our procedure. Moreover, Ghosal and van der Vaart (2001)'s result on the convergence rate of the Dirichlet mixture of normal model imply that the problem is nearly parametric, and therefore a fast convergence rate of our method is expected.

Chapter 4

Dirichlet Mixture Regression

In this chapter we show how to implement our proposed method to estimate the regression function under the mixture structure. It includes the specification of updating scheme of posterior using Gibbs sampling method and updating corresponding of hyperparameters. However, we do not actually update the hyperparameters; rather we replace by the empirical estimates. We do this to minimize computational complexity. Some theoretical results based on certain well known theorems about the Dirichlet mixture model will also be presented along the sequel.

4.1 Estimation via MCMC sampling

Suppose that we have the following model.

$$\boldsymbol{X_i}|\theta_i \overset{ind}{\sim} N(\theta_i, \boldsymbol{\Sigma}), \quad i = 1, \ldots, n$$

$$\theta_i|G \overset{iid}{\sim} G,$$

$$G \sim DP(M, G_0), \tag{4.1}$$

where $G_0(\theta) = \phi(\theta; \boldsymbol{m}, \boldsymbol{\Phi})$ is a p.d.f. of multivariate normal distribution with mean vector $\boldsymbol{m}$ and covariance matrix $\boldsymbol{\Phi}$.

Note here, the center measure is considered to have hyperparameters $\boldsymbol{m}$ and $\boldsymbol{\Phi}$. In this way we can expect a more flexible model compared to using a fixed center measure. The $p-$variate case is considered here since it includes the univariate case.

Then,

$$\theta_i|\theta_{-i}, \boldsymbol{X} \sim q_{0,i} G_i(\theta_i) + \sum q_{j,i} \delta_{\theta_j}(\theta_i), \tag{4.2}$$

where

$$q_{j,i} \propto \phi(\boldsymbol{X_i}; \theta_i, \boldsymbol{\Sigma}),$$

$$q_{0,i} \propto M \int \phi(\boldsymbol{X_i}; \theta_i, \boldsymbol{\Sigma}) dG_0(\theta),$$

$$\sum_{j \neq i} q_{j,i} + q_{0,i} = 1,$$

and $G_i(\theta_i)$ is the posterior distribution of θ_i given $\boldsymbol{X_i}$ with the prior $G_0(\theta_i)$ and the likelihood $N(\theta_i, \boldsymbol{\Sigma})$. That is,

$$dG_i(\theta_i) \propto \phi(\theta_i; \boldsymbol{X_i}, \boldsymbol{\Sigma}) \phi(\theta_i; \boldsymbol{m}, \boldsymbol{\Phi}).$$

There is also another equivalent form of (4.2).

$$\theta_i|\theta_{-i}, \boldsymbol{X} \sim q_{0,i} G_i(\theta_i) + \sum_{\phi_k \in \phi_{-i}} n_{-i,k} q_{k,i} \delta_{\phi_k}(\theta_i).$$

Given the assumptions in (4.1), the distribution of s_i given s_{-i}, θ^*_{-i} and $\boldsymbol{X}$ is the following:

Let θ^* be the vector of the distict value of θ. For $j = 1, \ldots, k_{-i}$,

$$\begin{aligned} P(s_i = j | s_{-i}, \theta^*_{-i}, \boldsymbol{X}) &\propto n_{-i,j}\phi(\boldsymbol{X_i}; \theta^*_j, \boldsymbol{\Sigma}), \\ P(s_i = k_{-i}+1 | s_{-i}, \theta^*_{-i}, \boldsymbol{X}) &\propto M \int \phi(\boldsymbol{X_i}; \theta^*, \boldsymbol{\Sigma}) dG_0(\theta^*). \end{aligned} \tag{4.3}$$

Since it is useful to consider only the update of s rather than s and θ^* both, we can integrate out θ^*_j from the above theorem.

Algorithm (3)

Given the assumptions in (4.1), the distribution of s_i given s_{-i} and $\boldsymbol{X}$ is the following: For $j = 1, \ldots, k_{-i}$,

$$\begin{aligned} P(s_i = j | s_{-i}, \boldsymbol{X}) &\propto n_{-i,j} \int \phi(\boldsymbol{X_i}; \theta^*, \boldsymbol{\Sigma}) dH_{-i,j}(\theta^*), \\ P(s_i = k_{-i}+1 | s_{-i}, \boldsymbol{X}) &\propto M \int \phi(\boldsymbol{X_i}; \theta^*, \boldsymbol{\Sigma}) dG_0(\theta^*) \\ &\propto M\phi(\boldsymbol{X_i}; \boldsymbol{m}, \boldsymbol{\Sigma} + \boldsymbol{\Phi}), \end{aligned} \tag{4.4}$$

where $H_{-i,j}$ is the posterior distribution based on the prior G_0 and all observations $\boldsymbol{X_l}$ for which $l \in I_j$ and $l \neq i$. That is,

$$dH_{-i,j}(\phi) = \frac{1}{C_{i,j}} \Big[\prod_{l \in I_j, l \neq i} \phi(\boldsymbol{X_l}; \theta^*, \boldsymbol{\Sigma}) \Big] dG_0(\theta^*),$$

where $C_{i,j}$ is the normalizing constant.

Note that here, the above formula can be simplified by using certain updating equation. The first quantity can be expressed as the following:

$$\int \phi(\boldsymbol{X_i}; \theta^*, \boldsymbol{\Sigma}) dH_{-i,j}(\theta^*) = \frac{\int \phi(\boldsymbol{X_i}; \theta^*, \boldsymbol{\Sigma}) \Big[\prod_{l \in I_j, l \neq i} \phi(\boldsymbol{X_l}; \theta^*, \boldsymbol{\Sigma}) \Big] \phi(\theta^*; \boldsymbol{m}, \boldsymbol{\Phi}) d\theta^*}{\int \Big[\prod_{l \in I_j, l \neq i} \phi(\boldsymbol{X_l}; \theta^*, \boldsymbol{\Sigma}) \Big] \phi(\theta^*; \boldsymbol{m}, \boldsymbol{\Phi}) d\theta^*}. \tag{4.5}$$

The only difference between the numerator and denominator is in the term $\phi(\boldsymbol{X_i}; \theta_j^*, \boldsymbol{\Sigma})$ which is additional in the numerator. When we add or delete one datum from the original, we can easily update the sample mean and sample variance without recomputing the same for the whole data except the deleted observation. This should be helpful in calculating above ratio of two integrals. The result is again the ratio of two normal p.d.f.'s which is multiplied with another normal p.d.f. The equation of (4.5) can be written as

$$\frac{n_{-i,j}}{n_{-i,j}+1}\phi(\boldsymbol{X_i}; \overline{\boldsymbol{X}}_{+i,j}, \frac{n_{-i,j}}{n_{-i,j}+1}\boldsymbol{\Sigma}) \times \frac{\phi(\boldsymbol{m}; \overline{\boldsymbol{X}}_{+i,j}, \frac{1}{n_{-i,j}+1}\boldsymbol{\Sigma}+\boldsymbol{\Phi})}{\phi(\boldsymbol{m}; \overline{\boldsymbol{X}}_{-i,j}, \frac{1}{n_{-i,j}}\boldsymbol{\Sigma}+\boldsymbol{\Phi})}, \tag{4.6}$$

where $\overline{\boldsymbol{X}}_{-i,j} = \frac{1}{n_{-i,j}} \sum_{l \in I_j, l \neq i} \boldsymbol{X_l}$ and $\overline{\boldsymbol{X}}_{+i,j} = \frac{1}{n_{-i,j}+1}\Big(\sum_{l \in I_j, l \neq i} \boldsymbol{X_l} + \boldsymbol{X_i} \Big)$.

4.1.1 Estimation of $\boldsymbol{\Sigma}$ and m, Φ

We assumed that $\boldsymbol{\Sigma}$ is the same with all $\boldsymbol{X_i}$, but the more general setting is possible as the following.

$$\boldsymbol{X_i}|\theta_i \overset{ind}{\sim} N(\theta_i, \boldsymbol{\Sigma}_i), \quad i = 1, \ldots, n, \tag{4.7}$$

$$\theta_i | G \overset{iid}{\sim} G,$$

$$G \sim DP(M, G_0).$$

In this case, G_0 a distribution of two variables, θ_i and $\boldsymbol{\Sigma}_i$. The rest of the procedure is straightforward. By obtaining the full conditional distributions of θ_i and $\boldsymbol{\Sigma}_i$, the posterior can be sampled from. Moreover, some prior for $(\boldsymbol{m}, \boldsymbol{\Phi})$, may be considered so that the fully Bayesian hierarchical inference is possible.

However, we can make it simpler so that the computing time can be substantially reduced. Instead of updating the hyperparameters $(\boldsymbol{m}, \boldsymbol{\Phi})$, we may replace them by their respective estimators:

$$\widehat{\boldsymbol{m}} = \overline{\boldsymbol{X}},$$

$$\widehat{\boldsymbol{\Phi}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^T - \widehat{\boldsymbol{\Sigma}}.$$

Also, it is necessary to estimate $\boldsymbol{\Sigma}$.

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{j=1}^{k}\sum_{l\in I_j}(\boldsymbol{X}_l - \overline{\boldsymbol{X}}_j)(\boldsymbol{X}_l - \overline{\boldsymbol{X}}_j)^T,$$

where $\overline{\boldsymbol{X}}_j = \frac{1}{n_j}\sum_{l\in I_j}\boldsymbol{X}_l$. The estimator $\widehat{\boldsymbol{\Phi}}$ is a positive definite matrix because

$$\begin{aligned}
\widehat{\boldsymbol{\Phi}} &= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^T - \frac{1}{n}\sum_{j=1}^{k}\sum_{l\in I_j}(\boldsymbol{X}_l - \overline{\boldsymbol{X}}_j)(\boldsymbol{X}_l - \overline{\boldsymbol{X}}_j)^T \\
&= \frac{1}{n}\sum_{j=1}^{k}\sum_{l\in I_j}\left[(\boldsymbol{X}_l - \overline{\boldsymbol{X}})(\boldsymbol{X}_l - \overline{\boldsymbol{X}})^T - (\boldsymbol{X}_l - \overline{\boldsymbol{X}}_j)(\boldsymbol{X}_l - \overline{\boldsymbol{X}}_j)^T\right] \\
&= \frac{1}{n}\sum_{j=1}^{k} n_j(\overline{\boldsymbol{X}}_j - \overline{\boldsymbol{X}})(\overline{\boldsymbol{X}}_j - \overline{\boldsymbol{X}})^T. \qquad (4.8)
\end{aligned}$$

The last term in (4.8) is essentially the standard estimator of covariance matrix in j^{th} group, so it is a positive definite matrix provided that $n_j > 1$. This assures that our estimator $\widehat{\boldsymbol{\Phi}}$ is a positive definite estimator.

4.1.2 The procedure of the estimation of $E(Y|X)$

We assumed that X observations are generated from several clusters and the cluster labels are not observable.

$$X|J = j \sim N(\theta_j, \boldsymbol{\Sigma}), \quad j = 1, \ldots, k,$$

where J is a latent variable which indicates the group membership. We assume that there are k distinct groups,

Note that here, X is a p-variate vector and $\mathbf{\Sigma}$ is $p \times p$ variance-covariance matrix. Within a specific group, the unknown mean function can be obtained,

$$Y|X = x, J = j \sim N(f_j(x), T^2), \quad j = 1, \ldots, k.$$

By the Bayes rule, we have,

$$P(J = j|X = x) = \pi_j(x) = \frac{\pi_j \phi_\Sigma(x - \theta_j)}{\sum_{l=1}^{K} \pi_l \phi_\Sigma(x - \theta_l)}, \tag{4.9}$$

where π_j is the prior distribution and $\phi_\Sigma(\cdot)$ is a multivariate normal p.d.f. with mean vector zero and variance-covariance matrix $\mathbf{\Sigma}$. Then eliminating J, we obtain

$$E(Y|X = x) = \sum_{j=1}^{k} E(Y|X = x, J = j) \times P(J = j|X = x) = \sum_{j=1}^{k} f_j(x)\pi_j(x). \tag{4.10}$$

Now, the suggested procedure to estimate mean function $E(Y|X = x)$ is the following.

(DP method - Ave)

- **Step 1:** Find the clustering only for X by using the Dirichlet mixture process prior in MCMC sampling.

- **Step 2:** Suppose we have $s^{(t)}$, the configuration vector, $t = 1, \ldots, N$, in the t^{th} MCMC step. Then we can fit the regression function within each group. Simple linear regression, the LASSO method or polynomial regression with quadratic or cubic may be considered. Let us say there are $\widehat{k}^{(t)}$ clusters. Then

$$\widehat{f}^{(t)}(x) = \sum_{j=1}^{\widehat{k}^{(t)}} \widehat{\pi}_j(x)\widehat{f}_j(x), \tag{4.11}$$

where $\widehat{\pi}_j(x) \propto \frac{n_j^{(t)}}{n}\phi_{\widehat{\mathbf{\Sigma}}}(x - \widehat{\phi}_j)$, $\widehat{\phi}_j = \frac{1}{n_j^{(t)}}\sum_{l \in I_j^{(t)}} X_l$, and $\widehat{f}_j$ is the estimated regression function in the j^{th} cluster.

- **Step 3:** Average out for all $\widehat{f}^{(t)}(x), \quad t = 1, \ldots, N,$

$$\widehat{f}(x) = \frac{1}{N}\sum_{t=1}^{N} \widehat{f}^{(t)}(x).$$

(DP method - Most likely)

Instead of averaging out regression fitted function with some weight as in (4.11), we may also consider the value of $\widehat{f}_j(x)$ corresponding to the j for which $\widehat{\pi}_j(x)$ is highest. This is also implemented in our simulations. We call it the "Most likely" method.

- **Step 1:** Find the clustering only for X by using DP prior in MCMC sampling.

- **Step 2:** Suppose we have $s^{(t)}$, the configuration vector, $t = 1, \ldots, N$, in the t^{th} MCMC step. Then we can fit the regression function within each group. Then

$$\widehat{f}^{(t)}(x) = \widehat{f}_{j^{*(t)}}(x),$$

where $j^{*(t)}$ is the index which indicate group with the highest probability attain, given x, that is, $j^{*(t)} = \text{argmax}_j \quad \pi_j(x)$. Note that $j^{*(t)}$ could have values $\{1, 2, \ldots, \widehat{k}^{(t)}\}$.

- **Step 3:** Average out for all $\widehat{f}^{(t)}(x), \quad t = 1, \ldots, N,$

$$\widehat{f}(x) = \frac{1}{N}\sum_{t=1}^{N} \widehat{f}^{(t)}(x).$$

4.1.3 Merging clusters

During the MCMC procedure, it may happen that some clusters contain only a few data points. Fitting the regression function in that cluster may be problematic. For example, there needs to be at least two points to fit the regression line in the univariate case. For higher dimensional cases, the problem could be more prominent. We suggest merging some clusters to a new cluster to avoid problems of small clusters.

Suppose we have k clusters, $I_1, \ldots, I_k$ where

$$I_j = \{i : s_i = j, \quad i = 1, \ldots, n\}.$$

Now, check whether $|I_j| \geq c$ for all $j = 1, \ldots, k$, where c is a fixed constant and $|I_j|$ is the number of elements in cluster I_j. If the condition holds, go to the next step in our procedure. If not, find j^* such that $|I_{j^*}| = \min_j |I_j|$. Then for all $s_i \in I_{j^*}$ rearrange

$$s_i = l, \quad l \neq j^*,$$

with the probability

$$P(s_i = l|X) = \pi_l(X_i) \quad \text{(see the equation (4.9))}.$$

Repeat this until above condition is satisfied.

4.2 Confidence band for regression function

In this section, we discuss a method of constructing a confidence band for a regression function. This band enables us to see the region in which the entire regression

function lies. Confidence limits in the standard regression problem is well-known. But, the difficulty in our model arises because the clusters are unknown.

Let x be a fixed point in the space of the regressor. Suppose we know $100(1-\alpha_0)\%$ confidence limits for a single mean response as L_j, U_j within cluster $j = 1, \ldots, k$, where L_j is the lower limit and U_j is the upper limit. Thus

$$P(L_j \leq E(Y|X = x) \leq U_j|J = j) = 1 - \alpha_0. \tag{4.12}$$

Our goal is to find $100(1 - \alpha)\%$ the confidence limit L and U such that

$$P(L \leq E(Y|X = x) \leq U) \geq 1 - \alpha. \tag{4.13}$$

We claim that if L is a $\frac{\gamma}{2}$ quantile of $L_1, \ldots, L_k$ and U is a $1-\frac{\gamma}{2}$ quantile of $U_1, \ldots, U_k$, respectively, where $\gamma = \alpha - \alpha_0$ and the probabilities are $\pi_1(x), \ldots, \pi_k(x)$, then above inequality (4.13) holds. Note that,

$$\begin{aligned}
P(E(Y|X = x) \leq L) &= \sum_{j=1}^{k} P(E(Y|X = x) \leq L|J = j)P(J = j) \\
&= \sum_{\{j:L_j \leq L\}} P(E(Y|X = x) \leq L|J = j)P(J = j) \\
&\quad + \sum_{\{j:L_j > L\}} P(E(Y|X = x) \leq L|J = j)P(J = j) \\
&\leq 1 \cdot \sum_{\{j:L_j \leq L\}} P(J = j) + \frac{\alpha_0}{2} \sum_{\{j:L_j > L\}} P(J = j) \\
&\leq \frac{\gamma}{2} + \frac{\alpha_0}{2}.
\end{aligned} \tag{4.14}$$

Similarly we can get, $P(E(Y|X) \geq U) \leq \frac{\gamma}{2} + \frac{\alpha_0}{2}$. Therefore, by letting $\alpha = \alpha_0 + \gamma$, we can get the inequality (4.13). This is a somewhat conservative confidence region. Within each group we choose $1 - \alpha_0$ confidence limit. Then our actual confidence band for the regression function is selected by the $\frac{\gamma}{2}$ quantiles of the confidence limits

along the groups. For example, let γ is given 0.025 and $\alpha_0 = 0.025$. Therefore if we want to find 95% confidence band, in each group at given MCMC step it is needed to obtain 97.5% confidence limits.

4.3 Theoretical Results

In this section we show weak posterior consistency for estimating the mixture density by generalizing the work of Ghosal *et al.* (1999) to the multivariate case. Suppose we have a Dirichlet mixture model as (4.1) where $\Sigma = \sigma^2 \mathrm{I}_p$. The sample, X_i is a p-variate vector and the p.d.f. of X_i is given as

$$f_{\sigma,P}(x) = \int N(x; \theta, \sigma^2) dP(\theta). \tag{4.15}$$

Note that this is the convolution $\phi_\Sigma * P$. The prior for (σ, P) is denoted by $\mu \times \Pi$.

Theorem 3 *Let f_0 be the true density of the form*

$$f_0(x) = f_{\sigma_0, P_0}(x) = \int \phi_{\Sigma_0}(x - \theta) dP_0(\theta),$$

where x, θ are p-variate vector and $\Sigma_0 = \sigma_0^2 I_p$. If P_0 is compactly supported and belongs to the support of Π and σ_0 is in the support of μ, then $\Pi(K_\varepsilon(f_0)) > 0$ for all $\varepsilon > 0$.

Proof. The true P_0 is a finite discrete distribution so we can assume that $P_0(A) = 1$ where $A = [-a, a] \times \cdots \times [-a, a]$. Since P_0 is in the weak support of Π then, $\Pi\{P : P(A) > \frac{1}{2}\} > 0$.

We want to show that

$$\Pi\Big\{f : \int f_0 \log\Big(\frac{f_{\Sigma_0,P_0}}{f_{\Sigma,P}}\Big) < \varepsilon\Big\} > 0. \tag{4.16}$$

We may split the expression in (4.16) into two parts.

$$\int f_0 \log\Big(\frac{f_{\Sigma_0,P_0}}{f_{\Sigma,P}}\Big) = \int f_0 \log\Big(\frac{f_{\Sigma_0,P_0}}{f_{\Sigma,P_0}}\Big) + \int f_0 \log\Big(\frac{f_{\Sigma,P_0}}{f_{\Sigma,P}}\Big). \tag{4.17}$$

First, we show that the second term on the right hand side of (4.17) is less than $\frac{\varepsilon}{2}$ with positive prior probability.

We need to divide the space $\mathbb{R}^p$ into 3^p pieces. Each part can be classified by three types.

$$B_1 = \prod_{i=1}^{p}[-b, b], \quad B_2 = \prod_{i=1}^{p}((-\infty, -b) \cup (b, \infty)), \quad B_3 = (B_1 \cup B_2)^c,$$

where b is chosen by the following. For $\eta > 0$, choose b such that

$$\int_{x \in B_1^c} \max(1, \sum_{i=1}^{p} |x_i|, \sum_{i=1}^{p} x_i^2) f_0(x) dx_1 \cdots dx_p < \eta.$$

Now,

$$\begin{aligned} \int_{\mathbb{R}^p} f_0(x) \log(\frac{f_{\Sigma,P_0}(x)}{f_{\Sigma,P}(x)}) dx &= \int_{B_1} f_0(x) \log\Big(\frac{f_{\Sigma,P_0}(x)}{f_{\Sigma,P}(x)}\Big) dx \\ &+ \int_{B_1^c} f_0(x) \log\Big(\frac{f_{\Sigma,P_0}(x)}{f_{\Sigma,P}(x)}\Big) dx, \end{aligned} \tag{4.18}$$

$$\begin{aligned}
\int_{B_1^c} f_0(x)\log\Big(\frac{f_{\Sigma,P_0}(x)}{f_{\Sigma,P}(x)}\Big)dx &\leq \int_{B_1^c} f_0(x)\log\Big(\frac{\int_A \phi_\Sigma(x-\theta)dP_0(\theta)}{\int_A \phi_\Sigma(x-\theta)dP(\theta)}\Big)dx \\
&= \int_{B_2} f_0(x)\log\Big(\frac{\int_A \phi_\Sigma(x-\theta)dP_0(\theta)}{\int_A \phi_\Sigma(x-\theta)dP(\theta)}\Big) \\
&\quad + \int_{B_3} f_0(x)\Big(\frac{\int_A \phi_\Sigma(x-\theta)dP_0(\theta)}{\int_A \phi_\Sigma(x-\theta)dP(\theta)}\Big)dx \\
&\leq \int_{B_2} f_0(x)\log\Big(\frac{\phi_\Sigma(x+a)}{\phi_\Sigma(x-a)P(A)}\Big)dx \\
&\quad + \int_{B_3} f_0(x)\Big(\frac{\int_A \phi_\Sigma(x-\theta)dP_0(\theta)}{\int_A \phi_\Sigma(x-\theta)dP(\theta)}\Big)dx \\
&< \Big(C_1\frac{1}{\sigma^2} + C_2\frac{2a}{\sigma^2} + \log 2\Big)\eta, \qquad (4.19)
\end{aligned}$$

provided that $P(A) > \frac{1}{2}$ and C_1, C_2 are the fixed constants depending only p.

To assist understanding the above inequality (4.19), it is helpful to consider the $\mathbb{R}^2$ case as an example. We have the partition of $\mathbb{R}^2$ as shown in Figure 4.1. In Figure 4.1, B_2 is the region (1) and B_3 is the region (2). First, consider the left bottom region $(-\infty, -b) \times (-\infty, -b)$ in B_2, then it is easy to show that in region B_2 the following inequality holds.

$$\begin{aligned}
\log\Big(\frac{\int_A \phi_\Sigma(x-\theta)dP_0(\theta)}{\int_A \phi_\Sigma(x-\theta)dP(\theta)}\Big) &\leq \log\Big(\frac{\phi_\sigma(x_1+a)\phi_\sigma(x_2+a)}{\phi_\sigma(x_1-a)\phi_\sigma(x_2-a)P(A)}\Big) \\
&\leq \frac{2a}{\sigma^2}(|x_1|+|x_2|) + \log 2. \qquad (4.20)
\end{aligned}$$

Similarly in all regions (1), above inequality (4.20) holds except changing signs in a. Now, consider B_3, region (2) in Figure 4.1. Suppose (x_1, x_2) is given in the middle bottom part, $[-b, b] \times (-\infty, -b)$, then also it can be shown by

$$\begin{aligned}
\log\Big(\frac{\int_A \phi_\Sigma(x-\theta)dP_0(\theta)}{\int_A \phi_\Sigma(x-\theta)dP(\theta)}\Big) &\leq \log\Big(\frac{\phi_\sigma(0)}{\phi_\sigma(a+b)}\frac{\phi_\sigma(x_2+a)}{\phi_\sigma(x_2-a)P(A)}\Big) \\
&\leq \frac{b^2}{\sigma^2} + \frac{2a|x_2|}{\sigma^2} + \log 2. \qquad (4.21)
\end{aligned}$$

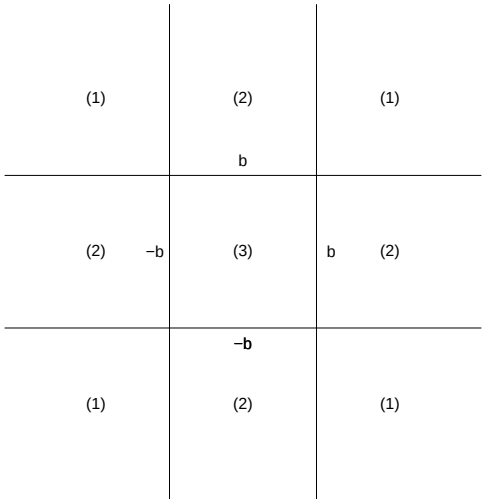


Figure 4.1: The partition of two dimensional space into $9(=3^2)$ pieces.

Note that,

$$b^2 \int_{-b}^{b} \int_{-\infty}^{-b} f_0(x) dx_2 dx_1 \leq \int_{-b}^{b} \int_{-\infty}^{-b} x_2^2 f_0(x) dx_2 dx_1.$$

In a similar way we can show that in all regions $B_1^c = B_2 \cup B_3$, either inequality (4.20) or (4.21) holds. Therefore, in region B_1^c the inequality (4.19) holds in two dimensional problem.

If we consider the compact set of B_1 then, $\phi_\Sigma(x-\theta)$ is bounded below for all $\theta \in \mathbb{R}^p$, so we can get

$$c = \inf_{|x| \in B_1} \inf_{\theta \in \mathbb{R}^p} \phi_\Sigma(x-\theta). \tag{4.22}$$

Note here $c > 0$.

The family of functions $\{\phi_\Sigma(x-\theta) : x \in B_1\}$ viewed as a set of functions of $\theta \in \mathbb{R}^p$, is uniformly equicontinuous. By the Arzela-Ascoli theorem, given any $\delta > 0$, there exist finitely many points $x^{(1)}, \ldots, x^{(m)}$ such that for any $x \in B_1$, there exists

an i with

$$\sup_{\theta\in\mathbb{R}^p} |\phi_\Sigma(x-\theta) - \phi_\Sigma(x^{(i)}-\theta)| < c\cdot\delta. \tag{4.23}$$

Let

$$N(P_0) = \Big\{P : \Big|\int \phi_\Sigma(x-\theta)dP_0(\theta) - \int \phi_\Sigma(x^{(i)}-\theta)dP(\theta)\Big| < c\delta, \quad i=1,\ldots,m\Big\}.$$

Then $N(P_0)$ is a weak neighborhood of P_0, $\Pi(B_p) > 0$.

Let $P \in N(P_0)$. Then for any $x \in B_1$, by choosing an appropriate $x^{(i)}$ from (4.23) and using a simple triangulation argument, we get

$$\Big|\int \phi_\Sigma(x-\theta)dP_0(\theta) - \int \phi_\Sigma(x-\theta)dP(\theta)\Big| < 3c\delta,$$

and $\int \phi_\Sigma(x-\theta)dP_0(\theta) > c\cdot P_0(A) = c$. Therefore,

$$\Big|\frac{\int \phi_\Sigma(x-\theta)dP(\theta)}{\int \phi_\Sigma(x-\theta)dP_0(\theta)} - 1\Big| < \frac{3c\delta}{c} = 3\delta.$$

By the fact that if $|x-1| < \epsilon$ then $|\frac{1}{x}-1| < \frac{\epsilon}{1-\epsilon}$ where $x > 0$, we have that

$$\Big|\frac{\int \phi_\Sigma(x-\theta)dP_0(\theta)}{\int \phi_\Sigma(x-\theta)dP(\theta)} - 1\Big| < \frac{3\delta}{1-3\delta}.$$

Then,

$$\int_{B_1} f_0(x)\log\Big(\frac{f_{\Sigma,P_0}(x)}{f_{\Sigma,P}(x)}\Big) < \int_{B_1} f_0(x)\Big|\frac{\int \phi_\Sigma(x-\theta)dP_0(\theta)}{\int \phi_\Sigma(x-\theta)dP(\theta)} - 1\Big| < \frac{3\delta}{1-3\delta}, \tag{4.24}$$

since $\log(x) < \log(1+|x-1|) < |x-1|$ for any $x > 0$.

Now, from (4.18), the second term can be expressed as

$$\int f_0(x)\log\Big(\frac{f_{\Sigma,P_0}(x)}{f_{\Sigma,P}(x)}\Big) < (3^p-1)\Big(\frac{2a}{\sigma^2} + \log 2\Big)\eta + \frac{3\delta}{1-3\delta}. \tag{4.25}$$

The first term in (4.18) decreases to 0 as $\Sigma \to \Sigma_0$ (i.e. $\sigma \to \sigma_0$) because of the following.

$$\frac{\int \phi_{\Sigma_0}(x-\theta)dP_0(\theta)}{\int \phi_{\Sigma}(x-\theta)dP_0(\theta)} < \sup_{\theta \in \mathbb{R}^p} \frac{\phi_{\Sigma_0}(x-\theta)}{\phi_{\Sigma}(x-\theta)}.$$

Since σ_0 is in the support of μ, for given any $\epsilon > 0$, choose a neighborhood $N(\sigma_0)$ of σ_0 such that if $\sigma \in N(\sigma_0)$, the first term on the right hand side of (4.18) is less than $\frac{\epsilon}{2}$. Then, choose η and δ so that for any $\sigma \in N$, the right hand side of (4.25) is less than $\frac{\epsilon}{2}$.

■

In order to prove the strong consistency, we need to apply Theorem 2 in Chapter 3 to the DM problem. There is the result in univariate case by Ghosal *et al* (1999), but the proof in multivariate case seems to be difficult at present. The difficulty lies in the estimate of entropy using their method. A more refined entropy estimate is expected to resolve the problem. If the support of base measure of Dirichlet process is a compact set the proof can be easily done along the extension of Theorem 7 in Ghosal *et al* (1999). In this case, the space of densities is compact with respect to L_1. Thus L_1 topology is equivalent to the weak topology. Arguably, the assumption of compact support is restrictive and in particular does not apply to the conjugate normal base measure.

Although we can not complete the verification of strong consistency of the DM in multivariate case, it is worth mentioning some results. Suppose we assume that f_n converges to f in L_1 sense. Then, by the well-known fact, it is true $P_n \xrightarrow{w} P$. Thus total variation convergence of densities implies weak convergence of mixing measures.

From now on we discuss the property of the number of mixture in our problem. According to Donoho (1988) it can be asserted that the k normal mixtures can not be approximated by less number of k mixture of normal. He considered the number of mixture complexity $K(F)$ is an integer-valued functional and it is only possible to make a one-sided nonparametric confidence statement.

Donoho (1988) showed that certain functionals, for example $K(F)$ and the number of modes of a density, are norm semi-continuous. For such functionals it is not possible to make two-sided nonparametric confidence statement but one-sided statement is possible. Using Dirichlet mixture in finding appropriate clusters will give us asymptotically at least the larger number of clusters than the true number of mixtures.

Definition 6 *The functional J is said to be norm lower semi-continuous if for every sequence F_n of distributions satisfying $\|F_n - F\| \to 0$, we have*

$$\liminf_{n\to\infty} J(F_n) \geq J(F),$$

where $\|\cdot\|$ is the Kolmogorov-Smirnov norm defined by

$$\|F - G\| = \sup_t |F(t) - G(t)|.$$

Let $\{G_\theta; \theta \in \Theta\}$ be a parameterized family of distributions and let $K(F)$ be the mixture complexity of F, that is the least number of components necessary to exactly represent F. Formally,

$$K(F) = \inf\{k : F = \sum_{i=1}^{k} \beta_i G_{\theta_i}\},$$

where $\sum \beta_i = 1, \beta_i \geq 0$ for all i.

Lemma 1 [**Donoho (1988)**] *The functional $K(F)$ is norm lower semicontinuous.*

The essential assumption of above result is that there exists a sequence F_n converging to F in Kolmogorov-Smirnov distance. Suppose under that assumption, we are in a certain MCMC step and we have the estimate of the number of clusters $\widehat{k}$. Then the event that $\widehat{k}$ is greater than or equal to the true number k_0 has high probability in an asymptotic sense. In fact, if there are more groups than the true one then the fitting within those groups does harm only with a higher variance. The number of observations falling in the true groups tends to infinity as the corresponding probability is positive. Thus the regression function within each group is consistently estimable provided that misclassification effects are small. On the other hand, merging two or more genuine groups by error introduces serious bias and hence underestimation of the true number of clusters is more harmful. In Figure 4.2 it is obvious that the mixture of three normals can approximate well if the true distribution is the mixture of two normals. But, the converse does not work well.

Further, since there are only finitely many groups, the number of observations corresponding to each group simultaneously go to infinity at the same rate of n no matter what the dimension p is. Thus our method clearly avoids curse of dimensionality. In this way we can hope that our estimator $\widehat{f}(x)$ has good asymptotic properties. Unfortunately Donoho's lower semi-continuity has been proved only under norm topology result, not in the weak sense. Moreover, we have not shown that effect of misgrouping is negligible. Therefore, we can not directly apply his result to our DM problem. We like to address this issue in the future work.

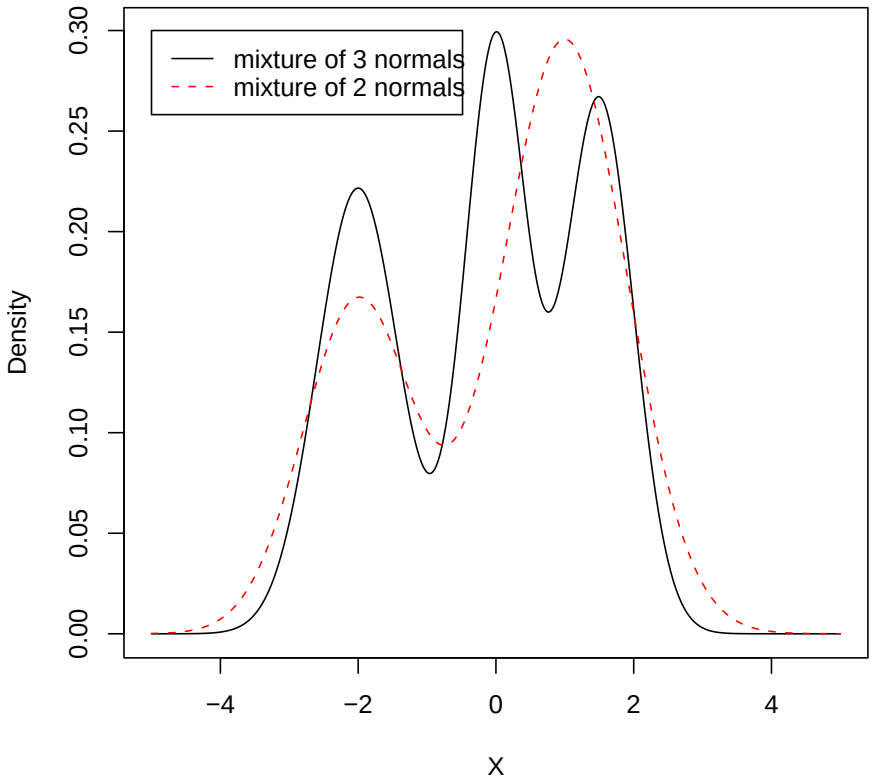


Figure 4.2: The plot of two and three mixture of normals.

Chapter 5

Simulation study and Applications

In this chapter we do several simulations to evaluate the performance of our method. First, data are generated from a mixture of normal distributions in the univariate and the multivariate case, and in different groups different regression functions are given with several parameters. We proceed to estimate the fitted regression function using our proposed method and also some other methods such as kernel and spline smoothing in the univariate case and MARS and GAM in the multivariate case. We compare the empirical L_1-error and L_2-error between estimates and the true regression function,

$$L_1E(\widehat{f}) = \frac{1}{n}\sum_{i=1}^{n}\Big|\widehat{f}(x_i) - f(x_i)\Big|, \tag{5.1}$$

$$L_2E(\widehat{f}) = \Big(\frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{f}(x_i) - f(x_i)\Big)^2\Big)^{\frac{1}{2}}.$$

Comparing the result to other methods will show the strength and weakness of our proposed method. Simulation studies are conducted under several different combinations of the true model and different sample sizes. To see how accurate the clustering

is, we use the measure of similarity which we call Rand's measure; see Section 3.2. We also apply our method to the challenging large p, small n problem. A simple simulation work is conducted by showing how it works. Some real data examples including microarray data for large p, small n problem are given in this chapter. All of simulation work is done by the program R (version 2.0.1.), a statistical programming language.

5.1 One dimension

To give a simple example we consider a univariate predictor X having the following distribution.

$$X \sim 0.3N(-0.9, \sigma^2) + 0.2N(-0.3, \sigma^2) + 0.4N(0.4, \sigma^2) + 0.1N(1.0, \sigma^2),$$

where $\sigma^2 = 0.01, 0.02, 0.03, 0.04$ are considered.

First generate 100 samples of X from above mixture of normals. And for generating Y four linear functions with different coefficients are given within each group.

$$f_1(x) = 2.3 + 1.1x, \quad f_2(x) = 1.5 - 0.6x,$$
$$f_3(x) = 0.8 + 0.9x, \quad f_4(x) = 1.7 - 0.2x,$$

with an additive noise term having mean 0 and variance $T^2 = 0.03$.

In the MCMC sampling step, we generate 5000 samples and ignore 1000 as a burn-in period. For appropriate selection of burn-in period, the trace plot of the number of distinct clusters was used. This simulation work is repeated 100 times and for one

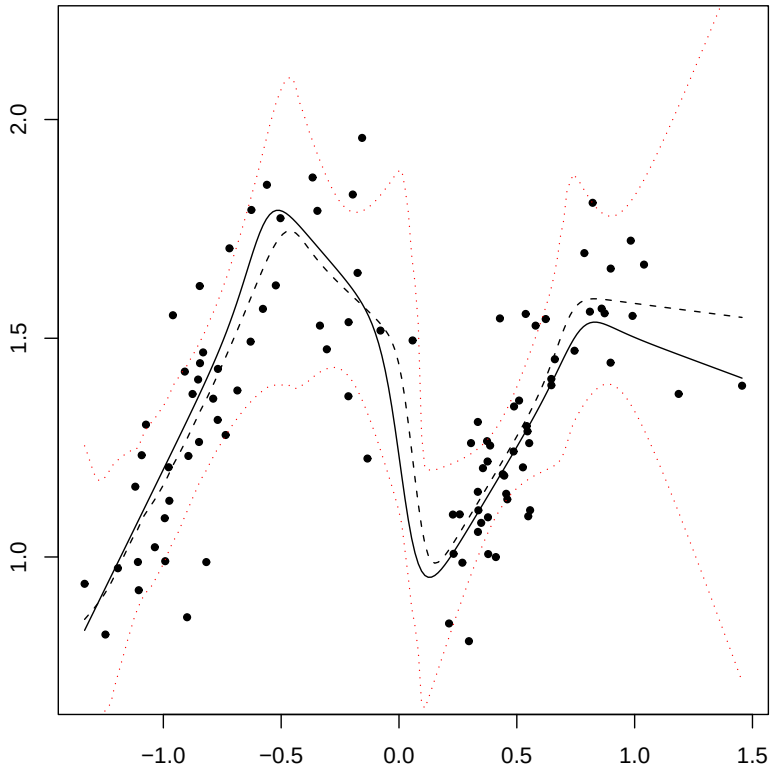
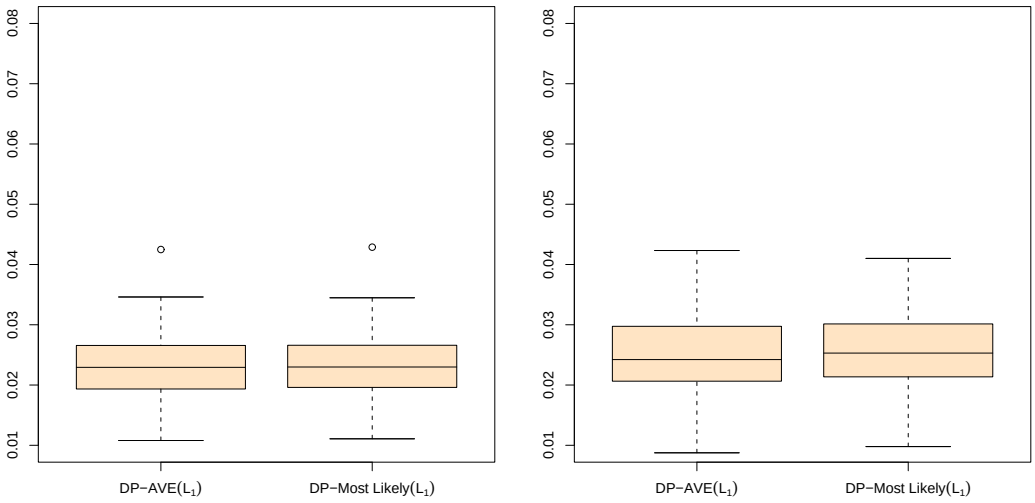


Figure 5.1: (One dimension) : The plots of data and fitted regression function with $\sigma^2 = 0.03$ and $T^2 = 0.03$. The solid line is the true mean function and dotted line is fitted line using DM-AVE. 95% confidence band is calculated as given in Section 4.2.

simulation it takes about 5 minutes. To compare with other classical methods, kernel method and spline smoothing are considered. We use the criterion of the empirical L_1-error and L_2-error. Rand's measure defined in Section 3.2 is computed at each MCMC steps. Since the appropriate cluster selection is crucial to our method, Rand's measure is monitored and the average value is given. Also the number of clusters in MCMC steps $\widehat{k}^{(t)}$ is also obtained.

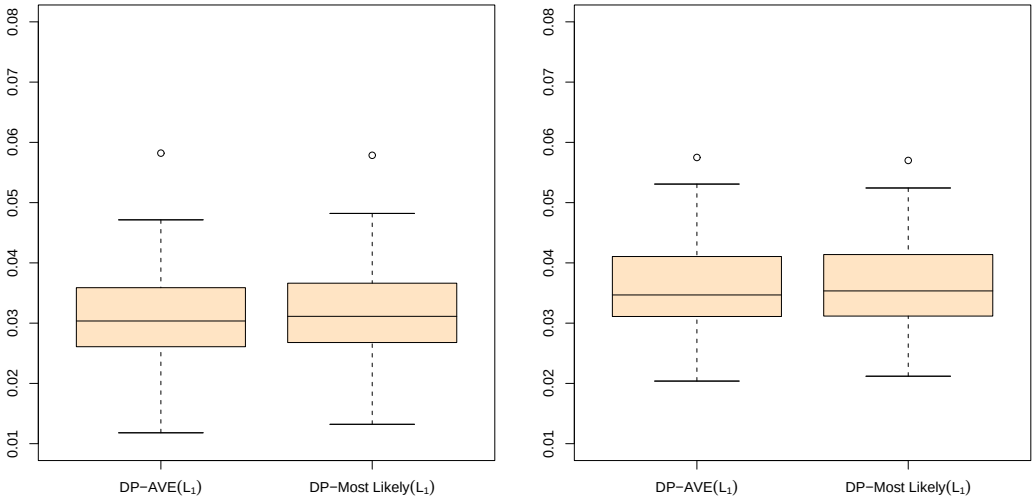
In Figure 5.2, our two suggested methods are considered. We have four different σ^2 values. The results are exactly same in this case except a few values so we only present one method, "DM-AVE". Figure 5.3 is the result for three methods including our method.

For kernel estimation, we used normal kernel regression estimate. This method was already implemented in R (version 2.0.1) program with package "KernSmooth" and function name is "ksmooth". There are a lot of bandwidth selection methods such as the 'plug-in bandwidth selection' suggested by Sheather and Jones (1991) and described in detail in Section 3.6 of Wand and Jones (1995). This can also be found in the package "KernSmooth" and "dpik" function. We measure the empirical L_2-error and L_1-error. It shows that our method has good performance for all the cases we investigated. The larger value of σ^2 means that each group are overlapped more than the small value case. Rand's measure decreases as σ^2 increases. This is expected because if some of clusters are severely overlapped, it is hard to determine the correct cluster. For the choice of σ^2, it is selected in the moderate level of overlapping data excluding two extreme cases. One is that there is only one cluster. The other is that the cluster is obvious to separate.



(a) for $\sigma^2 = 0.01$

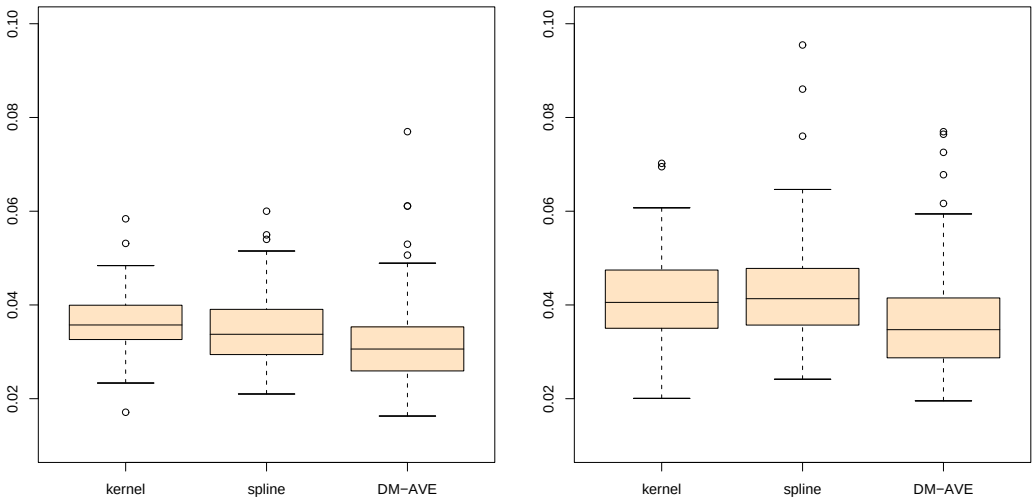
(b) for $\sigma^2 = 0.02$



(c) for $\sigma^2 = 0.03$

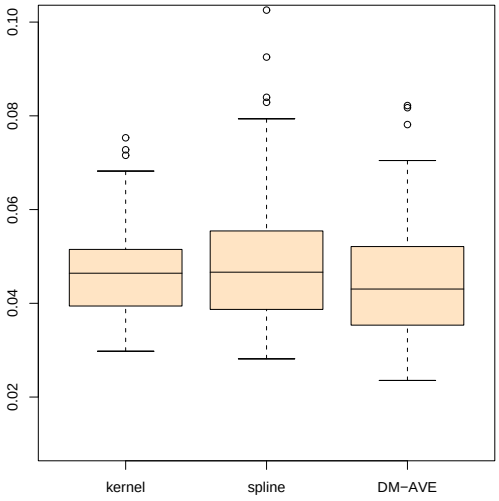
(d) for $\sigma^2 = 0.04$

Figure 5.2: (One dimension) : The plots of L_1-error with two Dirichlet mixture methods, "DM-AVE" and "DM-Most likely". Four different σ^2 are considered and the measures of similarity on average are 0.94, 0.87, 0.83 and 0.81, respectively.

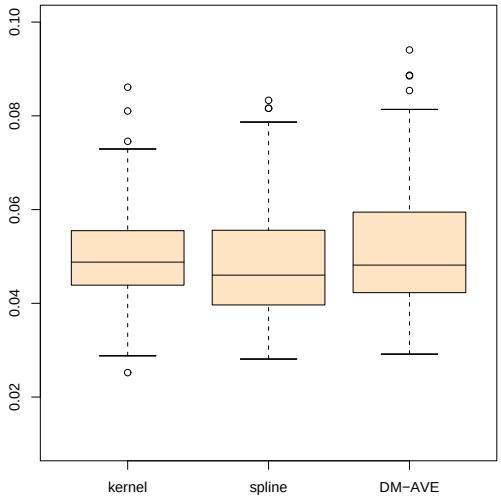


(a) for $\sigma^2 = 0.01$

(b) for $\sigma^2 = 0.02$



(c) for $\sigma^2 = 0.03$



(d) for $\sigma^2 = 0.04$

Figure 5.3: (One dimension) : The plots of L_2-error. Rand's measures on average are 0.94, 0.87, 0.83 and 0.81, respectively.

To compare above methods we use paired t-test. The result shows that DM-AVE is better than DM-ML in all cases.

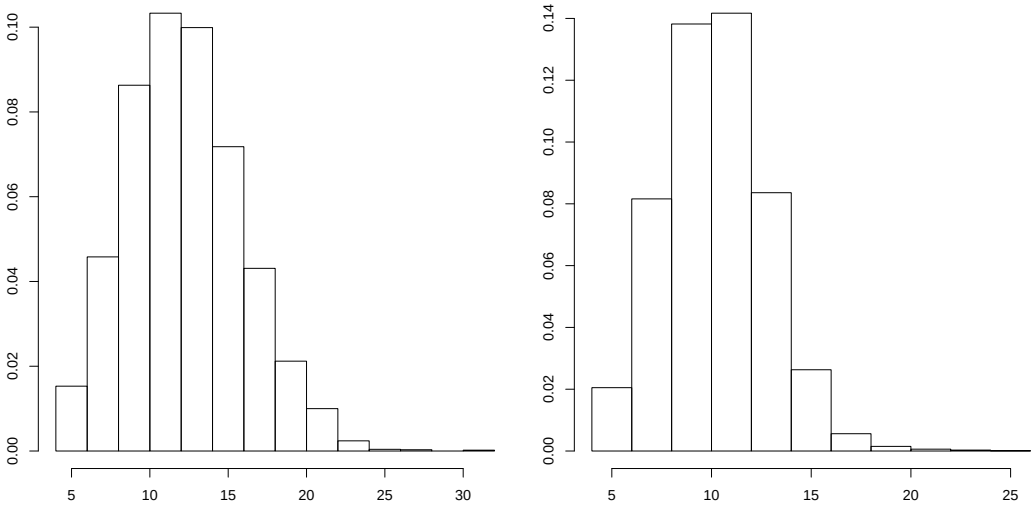
σ^2	0.01	0.02	0.03	0.04
DM-AVE vs DM-ML	< 0.01	< 0.01	< 0.01	< 0.01
DM-AVE vs kernel	< 0.01	< 0.01	0.02	0.88
DM-AVE vs spline	< 0.01	< 0.01	< 0.01	0.89

Table 5.1: The P-values from the paired t-test for the performances. The first row shows our suggested two method, DM-AVE and DM-ML are very similar to each other. The null hypothesis in the second row is that two means of L_2-error of kernel and DM-AVE methods are same. The third row is the result of spline smoothing and DM-AVE.

From Table 5.1, we can assure that in three cases, $\sigma^2 = 0.01, 0.02, 0.03$ with 95% significant level, the mean of DM-AVE method has average error smaller than other two methods. For the largest σ^2 value, we can not say that there exists significant difference among those methods.

The distribution of the number of clusters is given in Figure 5.4. As we mentioned that the estimated number of cluster tends to be greater than the true number asymptotically. We observe that all the values are greater than four, the true number of clusters. According to the plots in Figure 5.4, the larger σ^2 is, the less the number of clusters is. This can be investigated in higher dimension cases. We need to be careful in the interpretation of the number of clusters obtained from MCMC steps because it is not an estimate of the true number of clusters.

Let us consider the case that the true regression function is not linear. In the previous simulation, we assumed that the true regression function has the linear form and we used the linear regression method. But, here we want to look at the performance under misspecified models. Suppose X is generated by the following



(a) for $\sigma^2 = 0.01$ (b) for $\sigma^2 = 0.04$

Figure 5.4: (One dimension) : The plots of the distribution of k

distribution

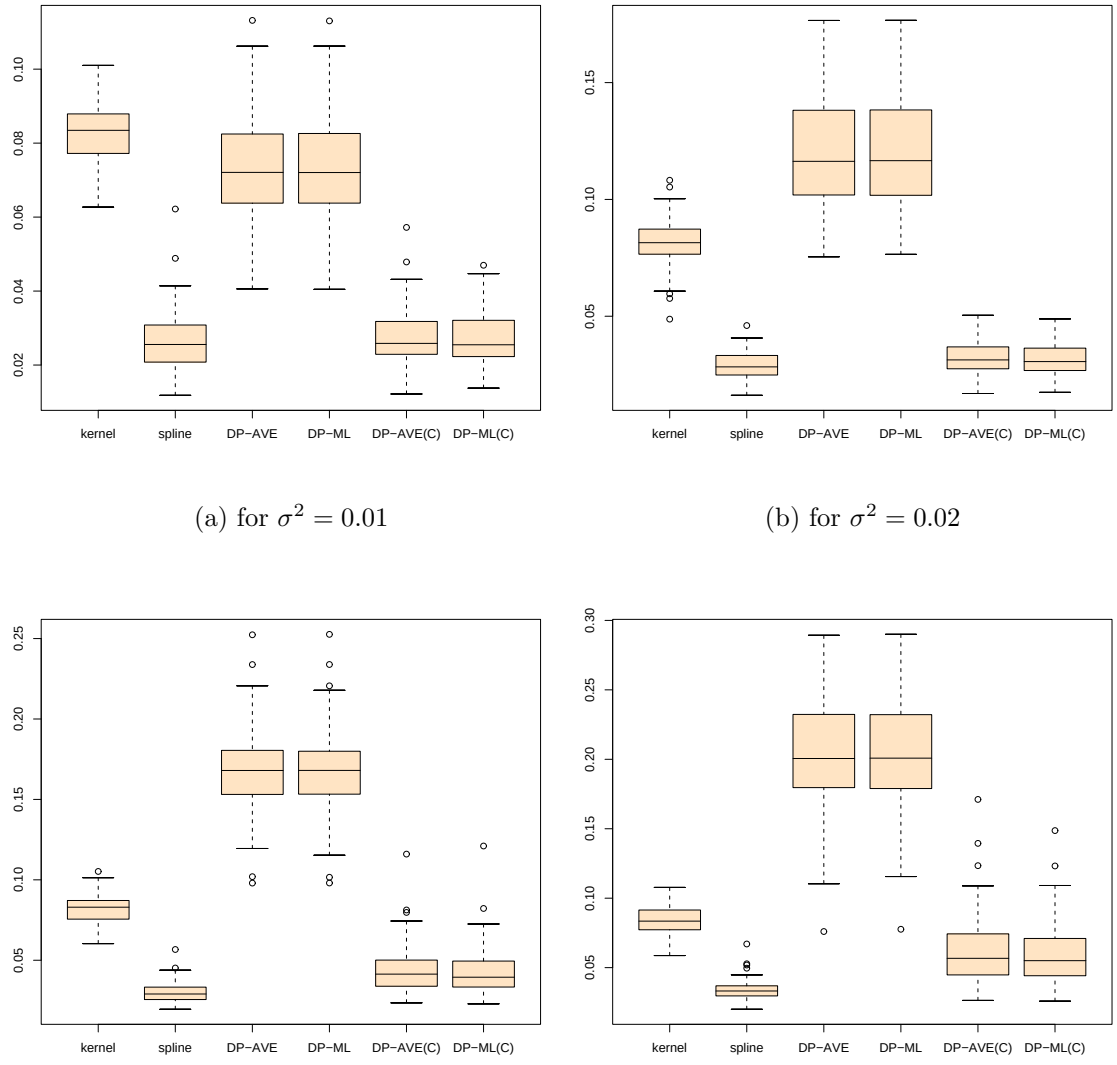
$$X \sim 0.3N(-1.9, \sigma^2) + 0.3N(-0.3, \sigma^2) + 0.4N(1.4, \sigma^2),$$

where $\sigma^2 = 0.01, 0.02, 0.03, 0.04$ are considered. The regression function in each cluster is given by

$$f_1(x) = 0.3 + 0.3\sin(2\pi x), \quad f_2(x) = 1.2 - 0.6\sin(2\pi x),$$
$$f_3(x) = 0.5 + 0.9\sin(2\pi x).$$

Note here that we have three clusters in this case. Other settings such as 5000 MCMC samples used and the number of repetition are same to previous ones. In this case, we use the cubic regression as well as the linear regression. Figure 5.5 tells us that the cubic regression performs better than the linear regression method and other

methods. Table 5.2 shows that DM-AVE method is better than kernel methods. The spline method seems to be best.



(a) for $\sigma^2 = 0.01$

(b) for $\sigma^2 = 0.02$

(c) for $\sigma^2 = 0.03$

(d) for $\sigma^2 = 0.04$

Figure 5.5: (One dimension) : The case where the true model is not linear. The plots of L_1-error are shown. Rand's measures are 0.97, 0.96, 0.95 and 0.94, respectively.

σ^2	0.01	0.02	0.03	0.04
DM-AVE(C) vs DM-ML(C)	0.99	0.98	1.00	0.97
DM-AVE(C) vs kernel	< 0.01	< 0.01	< 0.01	< 0.01
DM-AVE(C) vs spline	0.90	1.00	1.00	1.00

Table 5.2: The P-values from the paired t-test for the performances under misspecification. In our methods, DM-AVE and DM-ML, the cubic regression was used. The first row shows our suggested two method, DM-AVE and DM-ML are very similar to each other. The null hypothesis in the second row is that two means of L_1-error of kernel and DM-AVE methods are same. The third row is the result of spline smoothing and DM-AVE.

5.2 Two dimension

We first consider the two dimensional case, which is the simplest multivariate problem. Here, $\boldsymbol{X}$ is distributed as a mixture of bivariate normal. We set up our conditions very similar to the univariate case.

$$\boldsymbol{X} \sim 0.3N_2\Big(\begin{pmatrix}1\\1\end{pmatrix}, \Sigma\Big) + 0.2N_2\Big(\begin{pmatrix}2\\5\end{pmatrix}, \Sigma\Big) + 0.4N_2\Big(\begin{pmatrix}4\\0.5\end{pmatrix}, \Sigma\Big) + 0.1N_2\Big(\begin{pmatrix}5\\3\end{pmatrix}, \Sigma\Big)$$

$$E(Y|\boldsymbol{X}, J = j) = f_j(\boldsymbol{X}), \quad \text{for } j = 1, 2, 3, 4,$$

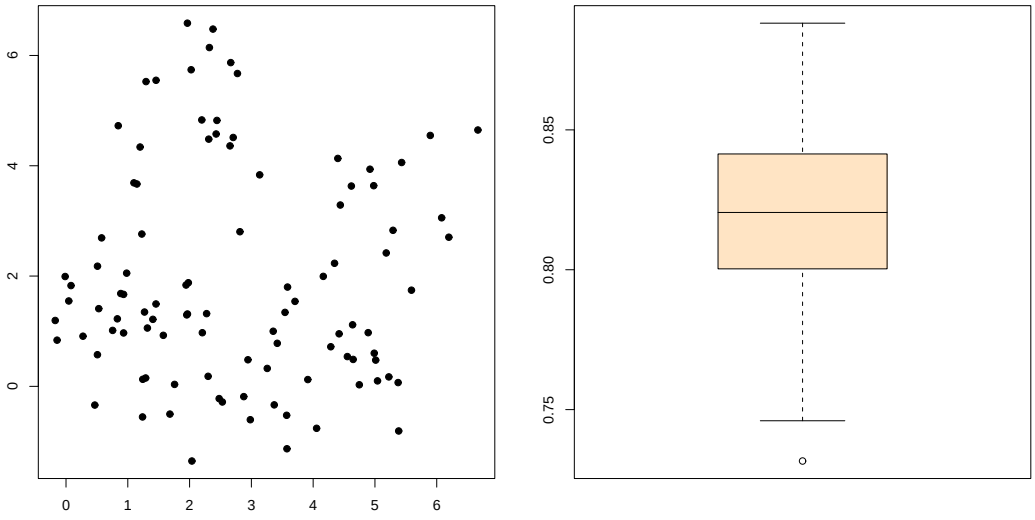
where

$$f_1(\boldsymbol{X}) = 2.3 + 1.1X_1 - 2X_2, \quad f_2(\boldsymbol{X}) = 1.5 - 0.6X_1 + 1.2X_2,$$

$$f_3(\boldsymbol{X}) = 0.8 + 0.9X_1 - 1.1X_2, \quad f_4(\boldsymbol{X}) = 1.7 - 0.2X_1 + 1.2X_2,$$

and $\Sigma = \sigma^2 I_2$, $\sigma^2 = 0.2, 0.4, 0.6, 0.8, 1.0$.

Figure 5.6 is the data plot of $\boldsymbol{X}$. The true number of clusters is four and those clusters are overlapped with each other. By the fact that σ^2 value determines how much the clusters are overlapped, we consider moderate values of σ^2. The extreme two cases are not of our interest. The variance, T^2 is still the same as the previous one, 0.1. Figure 5.6 shows four clusters are overlapped with each other.



(a) The plot of $\boldsymbol{X}$ (b) Boxplot of Rand's measure

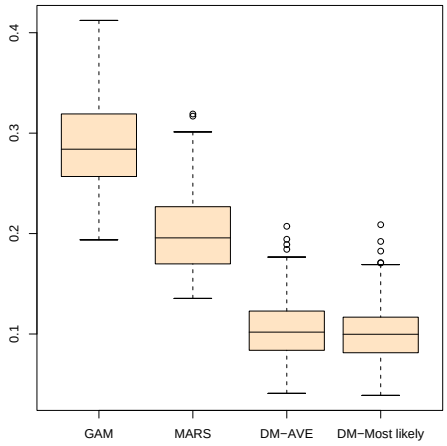
Figure 5.6: (Two dimension): data plot and Rand's measure with $\sigma^2 = 0.8$.

In MCMC sampling step, we generate 5000 samples and ignore 1000 as a burn-in period. To compare other methods such as GAM and MARS, we consider the L_2-error. To allow certain flexibility, we set the degree of interaction to two, which means that it allows interaction between two variables.

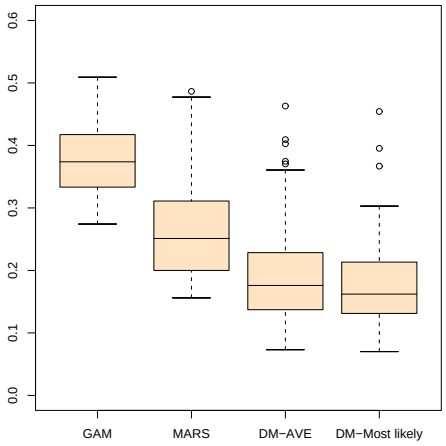
σ^2	GAM	MARS	DM-ML
0.2	< 0.01	< 0.01	0.78
0.4	< 0.01	< 0.01	1.00
0.6	< 0.01	< 0.21	0.99
0.8	< 0.01	0.01	1.00
1.0	< 0.01	0.29	1.00

Table 5.3: P-values from the paired t-test for the performances. At each row, different five σ^2 values are used. At each time, the comparison of DM-AVE and other three methods is conducted. In all cases, L_1-error was calculated and compared.

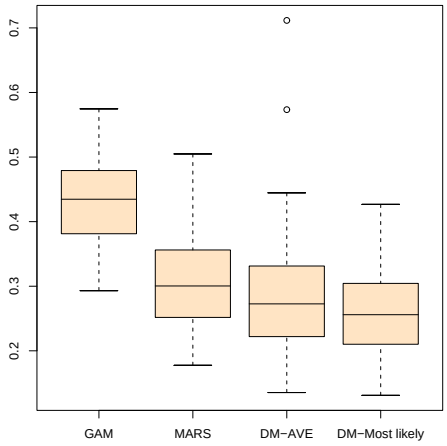
Table 5.3 gives the P-values of the comparison of the L_1-errors between GAM,



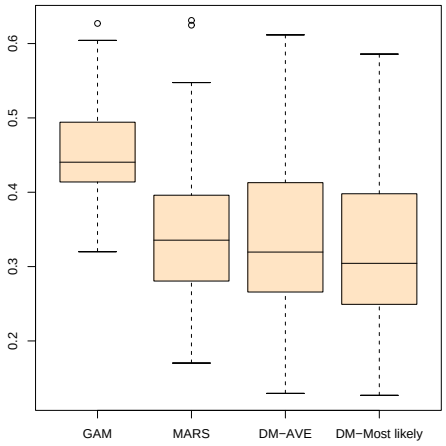
(a) for $\sigma^2 = 0.4$



(b) for $\sigma^2 = 0.6$



(c) for $\sigma^2 = 0.8$



(d) for $\sigma^2 = 1.0$

Figure 5.7: (Two dimension) : The plots of the L_2-error.

MARS, DM-ML and DM-AVE. The alternative hypothesis is that the mean L_1-error of DM-AVE is less than others. Therefore, with 95% confidence level we can say DM-AVE has lower L_1-error than GAM and MARS except in the $\sigma^2 = 1.0$ case. Again, there is no significant difference between DM-AVE and DM-ML. It is worth mentioning that we also calculated L_2-error and the result is very similar with L_1-error case.

5.3 Higher dimension

We consider the case when the predictor $\boldsymbol{X}$ has 10 variables and distributed as a mixture of four multivariate normal:

$$\boldsymbol{X} \sim \sum_{j=1}^{4} \omega_j N_{10}\Big(\boldsymbol{\mu}_j, \Sigma\Big),$$

$$E(Y|\boldsymbol{X}, J = j) = f_j(X_1, \ldots, X_{10}), \quad \text{for } j = 1, 2, 3, 4.$$

We let $\Sigma = \sigma^2 I_{10}$, $\sigma^2 = 0.2, 0.3, 0.4, 0.5$ and $T^2 = 0.1$. The four mean vector $\boldsymbol{m}_j$ is chosen as the following:

$$\begin{aligned} \boldsymbol{\mu}_1 &= (1,1,1,1,2,4,2,4,4,0.5)^T, \quad \boldsymbol{\mu}_2 = (4,0.5,5,3,5,3,1,1,1,1)^T, \\ \boldsymbol{\mu}_3 &= (2,5,2,5,4,0.5,4,0.5,5,3)^T, \quad \boldsymbol{\mu}_4 = (5,3,1,1,1,1,2,5,2,5)^T. \end{aligned}$$

Let the weights be $\omega = (0.3, 0.2, 0.4, 0.1)$. We generated 200 sample data from the following formula:

$$f_j(x) = \alpha_j + \boldsymbol{\beta}_j^T x, \quad j = 1, \ldots, 4,$$

where $(\alpha_1, \ldots, \alpha_4) = (2.3, 1.5, 0.8, 1.7)$ and $\boldsymbol{\beta}_j$'s are the vectors of length 10 given by

σ^2	0.2	0.3	0.4	0.5
$\boldsymbol{\beta}$	(1)	(2)	(3)	(4)
$\boldsymbol{\beta}^*$	(5)	(6)	(7)	(8)
$\boldsymbol{\beta}^*$, optimal for LASSO s	(9)	(10)	(11)	(12)

Table 5.4: (Ten dimension): The number represents the simulation sequence. s is the tuning parameter in the LASSO regression scaled 0 to 1.

the following:

$$\begin{aligned}
\boldsymbol{\beta}_1 &= (1.1, -2, 1.1, -2, -0.6, 1.2, -0.6, 1.2, 0.9, -1.1)^T, \\
\boldsymbol{\beta}_2 &= (0.9, -1.1, -0.2, 1.2, -0.2, 1.2, 1.1, -2, 1.1, -2)^T, \\
\boldsymbol{\beta}_3 &= (-0.6, 1.2, -0.6, 1.2, 0.9, -1.1, 0.9, -1.1, -0.2, 1.2)^T, \\
\boldsymbol{\beta}_4 &= (-0.2, 1.2, 1.1, -2, 1.1, -2, -0.6, 1.2, -0.6, 1.2)^T. \qquad (5.2)
\end{aligned}$$

We consider another simulation when $\boldsymbol{\beta}_j$'s are the same except that the last four components are replaced by zero. Let us call these $\boldsymbol{\beta}_j^*$ for $j = 1, \ldots, 4$.

$$\begin{aligned}
\boldsymbol{\beta}_1^* &= (1.1, -2, 1.1, -2, -0.6, 1.2, 0, 0, 0, 0)^T, \\
\boldsymbol{\beta}_2^* &= (0.9, -1.1, -0.2, 1.2, -0.2, 1.2, 0, 0, 0, 0)^T, \\
\boldsymbol{\beta}_3^* &= (-0.6, 1.2, -0.6, 1.2, 0.9, -1.1, 0, 0, 0, 0)^T, \\
\boldsymbol{\beta}_4^* &= (-0.2, 1.2, 1.1, -2, 1.1, -2, 0, 0, 0, 0)^T. \qquad (5.3)
\end{aligned}$$

The simulation sequences are given in Table 5.4. We consider four different σ^2's and two models such as full β are given as in (5.2) or β^* where some of coefficients are zero. We consider the LASSO regression method with a given tuning parameter s or optimally selected based on the observations. Two fixed numbers for s, 0.5 and 0.7 are considered and the optimal value is chosen by using CV method.

In the MCMC sampling scheme, we generate 5000 samples after ignoring first

σ^2	0.2	0.3	0.4	0.5
vs GAM	< 0.01	< 0.01	0.02	0.94
vs MARS	< 0.01	< 0.01	< 0.01	0.66

Table 5.5: P-values from the paired t-test for the performances. The alternative hypothesis in the first row is that the mean of L_1-error of GAM is greater than that of DM-AVE method. The second row is the result of MARS versus DM-AVE.

σ^2	0.2	0.3	0.4	0.5
vs GAM	< 0.01	< 0.01	< 0.01	0.12
vs MARS	< 0.01	< 0.01	< 0.01	0.01

Table 5.6: P-values from the paired median test for the performances. The alternative hypothesis in the first row is that the median of L_1-error of GAM is greater than that of DM-AVE method. The second row is the result of MARS versus DM-AVE.

1000 as burn-in. Figure 5.8 shows that our method is still best in the first top panel among all others. Note that the LASSO regression method is used in DM-AVE. In other panels, the difference between DM and other two methods is smaller. In order to make a statistical decision, paired t-test is used again. Table 5.5 shows that in the cases with $\sigma^2 = 0.2, 0.3, 0.4$, our DM-AVE method outperforms other methods. However there is some evidence that the distribution of the mean of L_1-error is not distributed normally. Thus the nonparametric test, Wilcoxon's rank test will be used to test the median of the L_1-errors. According to Table 5.6, DM-AVE method is better than two other methods for almost all the four cases.

We consider sparse model, that is, there are some zero coefficients. In Figure 5.9 the result is very similar to previous ones. But, if we compare the P-value from paired t-test, Table 5.7 shows that the difference between DM method and other two method is greater than before. This can be explained by the fact that LASSO is an

effective method under the sparse model.

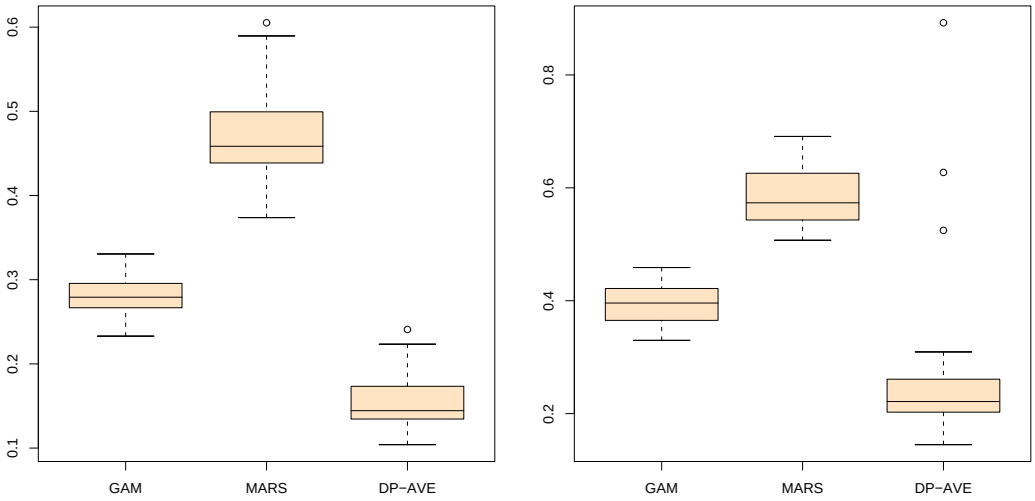
σ^2	0.2	0.3	0.4	0.5
DM-AVE vs GAM	< 0.01	0.28	0.62	0.81
DM-AVE vs MARS	< 0.01	0.16	0.58	0.88

Table 5.7: The P-values from the paired t-test for the performances. The alternative hypothesis in the first row is that the mean of L_2-error of GAM is greater than that of DM-AVE method. The second row is the result of MARS versus DM-AVE.

σ^2	0.2	0.3	0.4	0.5
vs GAM	< 0.01	< 0.01	< 0.01	< 0.01
vs MARS	< 0.01	< 0.01	< 0.01	0.02

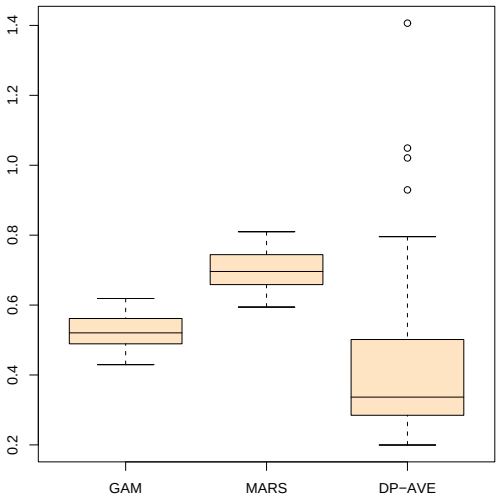
Table 5.8: The P-values from the paired t-test for the performances. The alternative hypothesis in the first row is that the mean of L_2-error of GAM is greater than that of DM-AVE method. The second row is the result of MARS versus DM-AVE.

To select the tuning parameter s, we used the common method of CV. Figure 5.10 shows the comparison of L_1-error with respect to the choice of s. We choose two fixed values as 0.5 and 0.7. The optimal choice of s among MCMC sampling will be various. However the performance of the choice of optimal s is better than other two fixed numbers. Note that the range of s value is $[0, 1]$ and if s is equal to 1, the LASSO coefficients are the same as the least squares estimates.

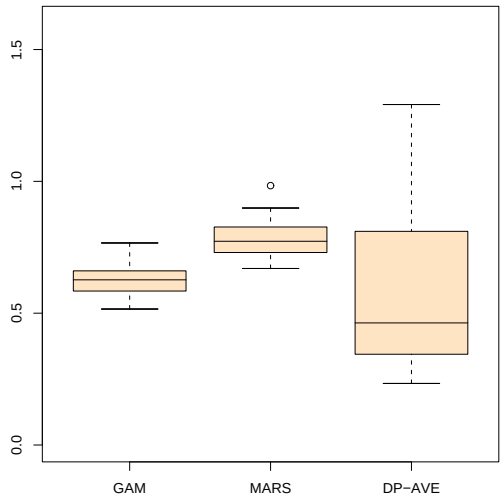


(a) for $\sigma^2 = 0.2$

(b) for $\sigma^2 = 0.3$

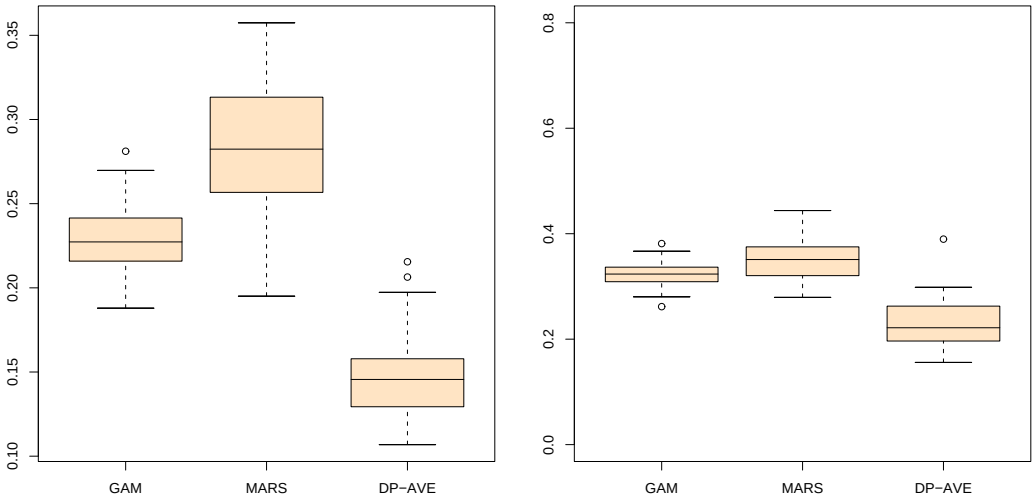


(c) for $\sigma^2 = 0.4$

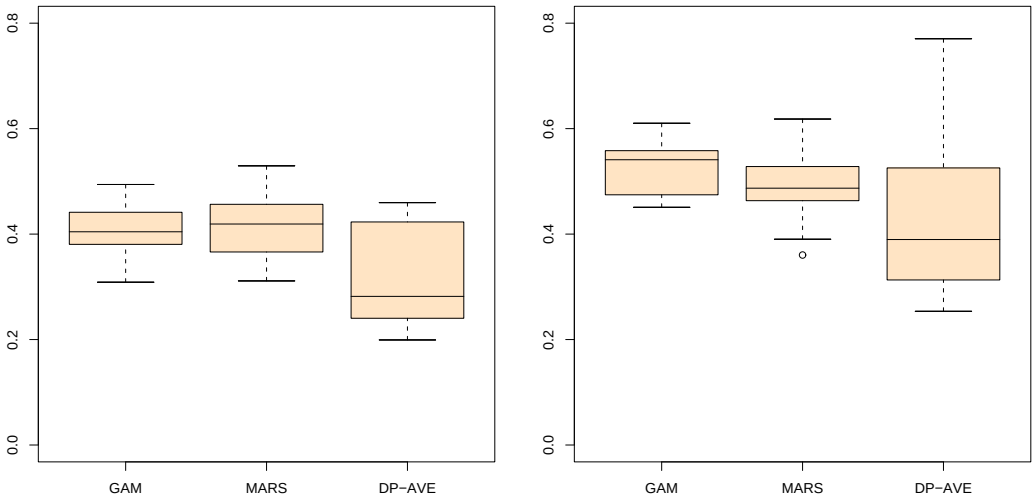


(d) for $\sigma^2 = 0.5$

Figure 5.8: (Ten dimension): The plots of L_1-error with respect to σ^2 values. Simulation numbers are (1) – (4) in the order. Average Rand's measures are 0.84, 0.77, 0.76 and 0.76, respectively.

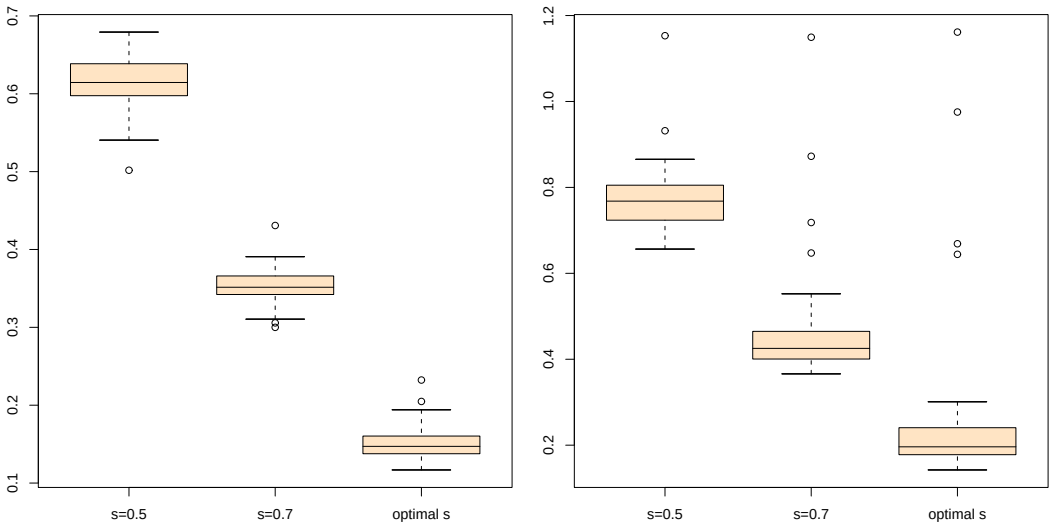


(a) for $\sigma^2 = 0.2$ (b) for $\sigma^2 = 0.3$



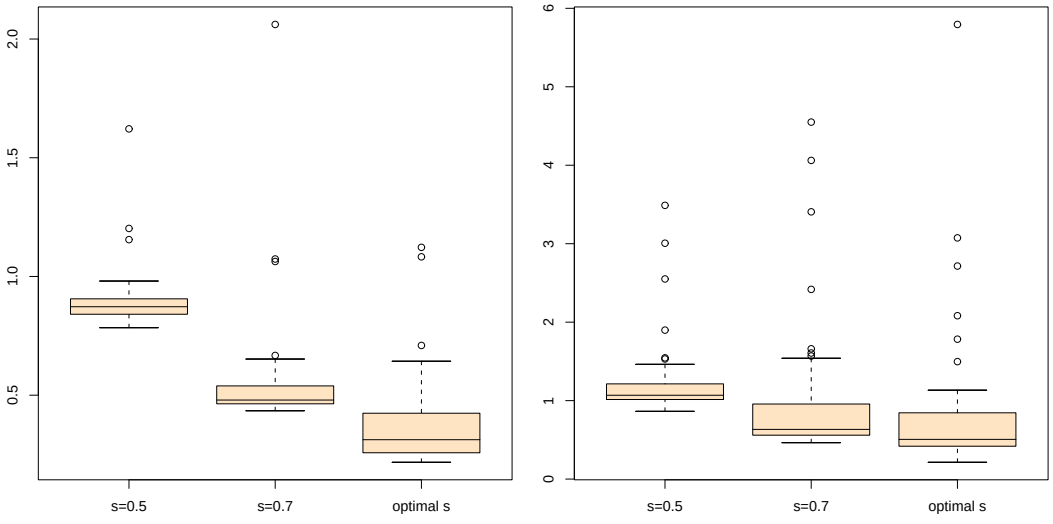
(c) for $\sigma^2 = 0.4$ (d) for $\sigma^2 = 0.5$

Figure 5.9: (Ten dimension): The plots of L_1-error with respect to σ^2 values. Simulation numbers are (5) – (8) in the order. Average Rand's measures are 0.84, 0.77, 0.76 and 0.76, respectively.



(a) for $\sigma^2 = 0.2$

(b) for $\sigma^2 = 0.3$



(c) for $\sigma^2 = 0.4$

(d) for $\sigma^2 = 0.5$

Figure 5.10: (Ten dimension): The plots of L_1-error with various s values, where s is the tuning parameter in LASSO.

5.4 Large p, small n case.

When the number of sample n is smaller than the number of variables p, the least square method does not apply anymore while LASSO automatically selects variables. However, its behavior tends to be erratic in that estimated coefficients can change sign. Also LASSO may fail to pick up all variables when a group is correlated. Because of these drawbacks of LASSO, Zou and Hastie (2005) suggested the so called method of "elastic net (EN)". To introduce EN criterion, we first give the definition of naive EN which is defined by finding $\boldsymbol{\beta}$ minimizing

$$\text{RSS } (\boldsymbol{\beta}, \lambda) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p} \beta_i^2. \tag{5.4}$$

The EN regression coefficient is nothing but a rescaled naive EN coefficient by the following,

$$\widehat{\boldsymbol{\beta}}_{\text{EN}} = (1 + \lambda_2)\widehat{\boldsymbol{\beta}}_{\text{naive EN}}.$$

Zou and Hastie (2005) argue that EN can select variables as in LASSO but does not suffer from similar difficulties. Their efficient algorithm called "LARS-EN" can solve the EN which is based on the proposed algorithm least angle regression (LARS) of Efron *et al.* (2004).

We want to use the Dirichlet mixture to find the clustering and EN to fit the regression within each cluster. The difficulty of using our method in this problem is that the covariance matrix can be singular. In equation (4.6), estimates of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Phi}$ are needed to calculate the inverse matrix. It is not possible to invert the resulting matrix because of its singularity. However, suppose we assume that $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, then we can substitute the estimate of σ^2, not of $\boldsymbol{\Sigma}$. This assumption is somewhat restrictive,

but we can work with this simple model so as to make the computation simple. Thus the parameter in the base measure, $\mathbf{\Phi} = \tau^2\mathbf{I}$ can be estimated accordingly. The estimates of those parameters are given by

$$\widehat{\sigma}^2 = \frac{1}{np}\sum_{r=1}^{p}\sum_{j=1}^{k}\sum_{i\in I_j}(\boldsymbol{X}_{i,r} - \overline{\boldsymbol{X}}_{j,r})^2 \tag{5.5}$$

and

$$\widehat{\tau}^2 = \frac{1}{np}\sum_{r=1}^{p}\sum_{i=1}^{n}(\boldsymbol{X}_{i,r} - \overline{\boldsymbol{X}}_{\bullet,r})^2 - \widehat{\sigma}^2, \tag{5.6}$$

where $\overline{\boldsymbol{X}}_{j,r} = \frac{1}{n_j}\sum_{i\in I_j}\boldsymbol{X}_{i,r}$ and $\overline{\boldsymbol{X}}_{\bullet,r} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_{i,r}$. The following simulation work will show how our method works in the large p, small n problem.

Let $p = 500$ and $n = 100$ be the number of dimension of $\boldsymbol{X}$ and samples, respectively.

$$\boldsymbol{X} \sim 0.3N_p(\mu_1, \mathbf{\Sigma}) + 0.7N_p(\mu_2, \mathbf{\Sigma}), \tag{5.7}$$

where μ_1 and μ_2 were chosen by the following

$$\mu_1 = (\underbrace{a_1, \ldots, a_{250}}_{250}, \underbrace{0.5, \ldots, 0.5}_{250}),$$

$$\mu_2 = (\underbrace{b_1, \ldots, b_{250}}_{250}, \underbrace{-0.5, \ldots, -0.5}_{250}),$$

where $a_i, b_i \overset{iid}{\sim} N(0, 0.5^2), i = 1, \ldots, 250$. Let $\mathbf{\Sigma}$ be $\sigma^2\mathbf{I}_p$, $\sigma^2 = 0.2^2$ and $T^2 = 0.5^2$. The linear functions in two groups are given by

$$f_1(\boldsymbol{X}) = \boldsymbol{X}\beta_1, \quad f_2(\boldsymbol{X}) = \boldsymbol{X}\beta_2, \tag{5.8}$$

where $\beta_1 = (\underbrace{1, \ldots, 1}_{30}, 0, \ldots, 0), \beta_2 = (\underbrace{-1, \ldots, -1}_{30}, 0, \ldots, 0)$. This data is artificial but it can at least illustrate the usefulness of our method in the large p, small n case.

We choose 1000 MCMC samples after 1000 burn-in. For selection of proper tuning parameters, we consider two dimensional CV method. For given choice of λ's (dropping subscript in λ_2) $(0.05, 0.1, 1, 5)$, the optimal s value is calculated. To evaluate how clustering is appropriate, the Rand's measure of similarity is used. The average Rand's measure is 0.62. The estimated L_2-error for four λ values are calculated. According to Table 5.9 the selection of $\lambda = 1$ is the best.

	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 1.0$	$\lambda = 5.0$
L_2-error	0.67	0.68	0.48	0.54

Table 5.9: (Large p, small n): L_2-error values for different λ values.

As a real data example, we consider *leukaemia* data consist of 200 genes and 38 samples analyzed in Zou and Hastie (2005). The original data consist of 7129 genes, but by prescreening 200 most significant genes are selected. There are two types of *leukaemia* – Type 1 *leukaemia* (acute lymphoblastic *leukaemia*) and type 2 *leukaemia* (acute myeloid *leukaemia*). There are 27 patients having Type 1 *leukaemia* and 11 patients having Type 2 *leukaemia*. To apply our method, we consider the Dirichlet mixture method to find two types of *leukaemia*. However, the result is somewhat undesirable because our method finds 19 clusters with only two points in each cluster. We may consider merging of some clusters in MCMC steps to rectify this problem.

Originally this is a classification problem where we already know which one is type 1 or 2. If we use the merging cluster method from Section 4.1.3 to the *leukaemia* data, 19 clusters can be reduced to the small number of clusters. Since all the clusters have two elements, we first randomly select among those clusters. For various constant c,

then 19 clusters reduced to a smaller number of clusters. To evaluate whether our clustering is correct or not, we can calculate Rand's measure. Before using merging method, Rand's measure was 0.43. This merging method improves the clustering because highest Rand's measure value is 0.66 after merging clusters when $c = 4$, see Figure 5.11.

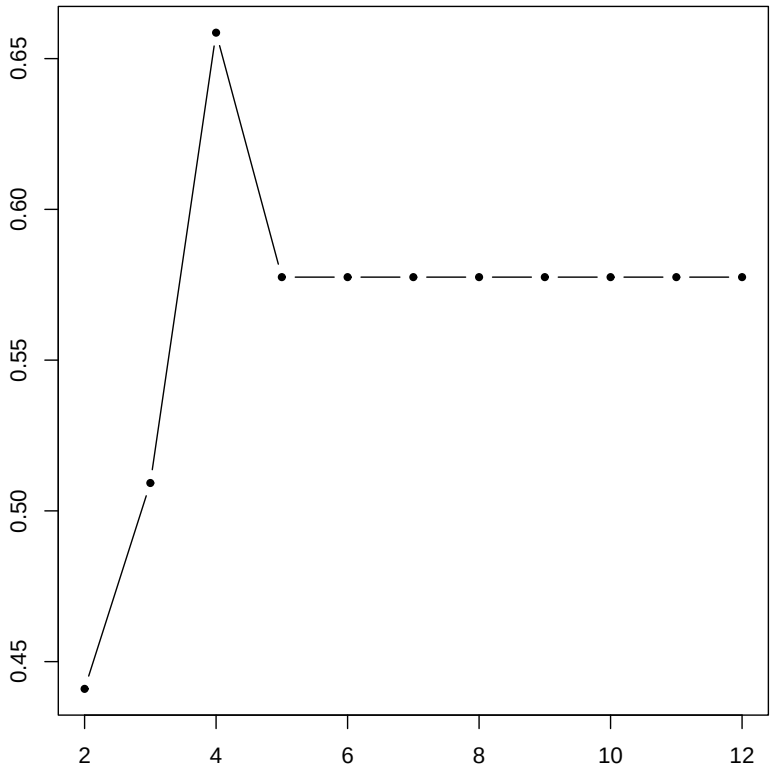


Figure 5.11: By using merging clusters Rand's measure is calculated for various choices of c.

5.5 Real data example

The data from a nonprofit organization that relies on fundraising campaigns to support their effort is considered as a real example. Originally this data is a sample data set in "Enterprise Miner" which is the data mining program from SAS. Dependent variable, Y measured how much the person actually donated in U.S. dollars and 8 predictors are chosen by,

$$
\begin{aligned}
X_1 &= \text{(donor's age)} , \\
X_2 &= \text{(income level (integer values 0 - 9))} , \\
X_3 &= \text{(\% of household in local government)} , \\
X_4 &= \text{(\% of household in state government)} , \\
X_5 &= \text{(\% of household in federal government)} , \\
X_6 &= \text{(total number of promotions)} , \\
X_7 &= \text{(donor's gifts to card promotions)} , \\
X_8 &= \text{(time between first and second donation)} , \\
X_9 &= \text{(donor's average gift).}
\end{aligned}
$$

Note that income level is an ordinal data but we can consider it as an interval variable. Our goal is to predict the donation for each selected person and investigate whether there are clusters. The data set has 200 samples and all eight variables are numeric values.

First, we use the LASSO method without clustering. By using optimal tuning parameter through five-fold CV method, only two variables, % of household in local

government and donor's average gift, have nonzero coefficients. This is a somewhat unexpected result. Now, we use the Dirichlet mixture method assuming that there are certain unknown clusters. To select the appropriate burn-in period, the trace plot of the number of clusters is monitored. Figure 5.12 shows 1000 burn-in is a reasonable choice. After burn-in, we get 5000 MCMC samples. The merging cluster method is used because there are some clusters having only a few points (two or three). Therefore, the number of clusters is decreased compared to previous one and shown in Figure 5.13.

For model construction in regression fitting, we need to transform Y, since there are some zero values. The new response variable $Y' = \log Y$ after adding a fixed small number, 0.1 to the zero observation. After the fitting procedure, the original variable Y will be calculated by the inverse transformation. Figure 5.14 shows the plots of two interesting variables, Age and Income level versus the amount of donation. The number of clusters is also of our interest. Figure 5.15 shows that the distribution of the modified number of clusters.

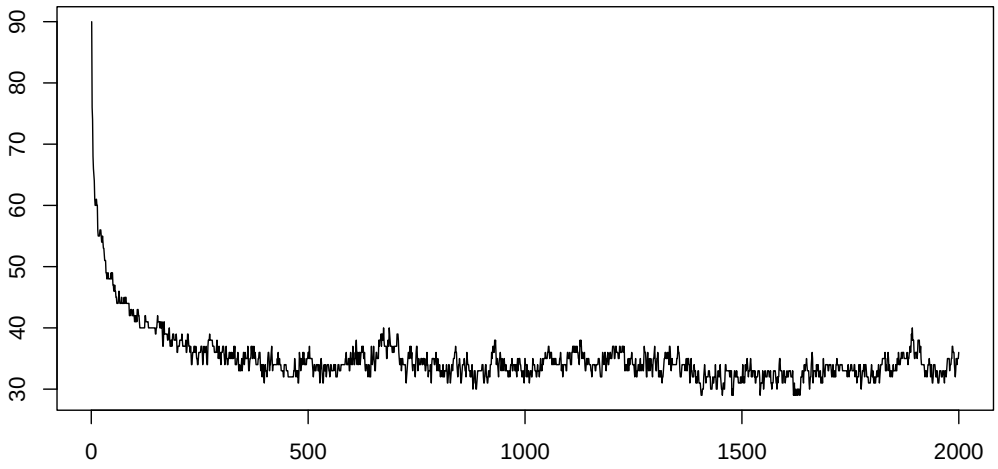


Figure 5.12: The trace plot of the number of clusters in first 2000 MCMC samples.

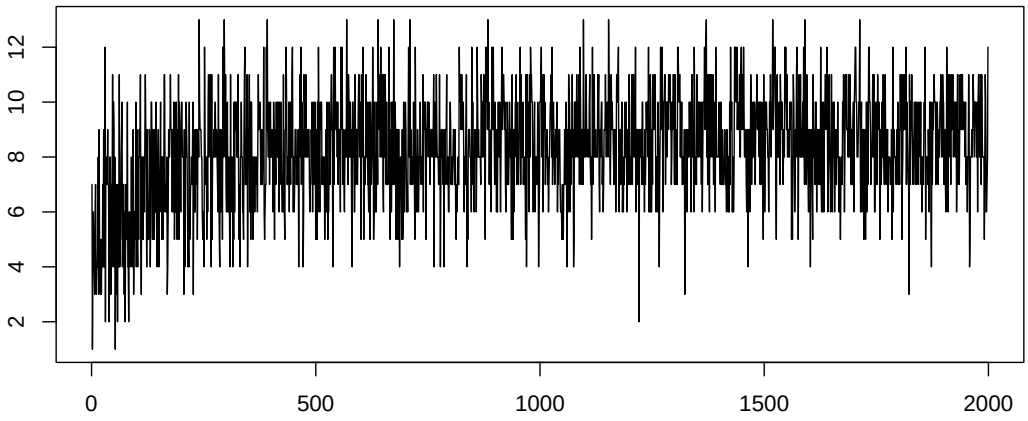
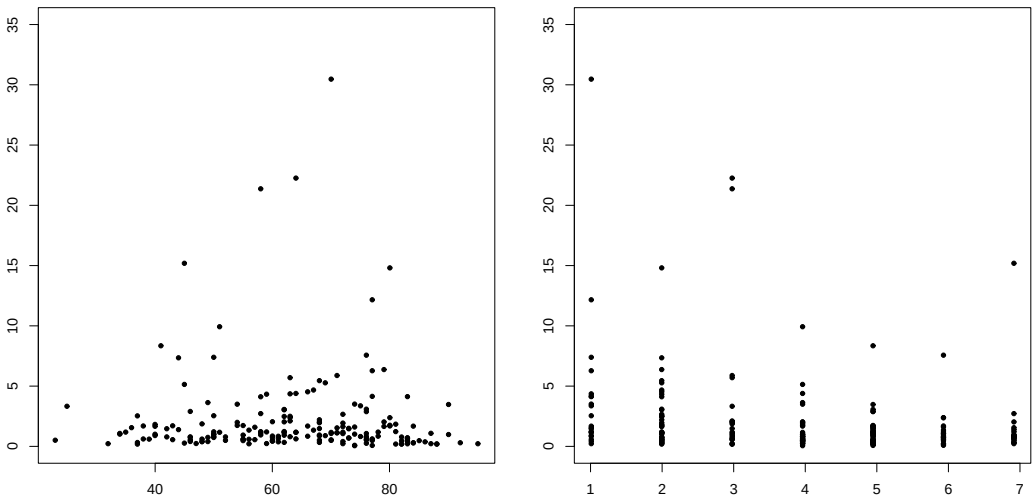


Figure 5.13: After using merging clusters the trace plot of the number of clusters in first 2000 MCMC samples.



(a) Age versus Donation

(b) Income versus Donation

Figure 5.14: The estimated values at given age and income level. Left panel shows that the amount of donation according to age. Right panel is the plot of donation and the income level ranging from 1 to 7.

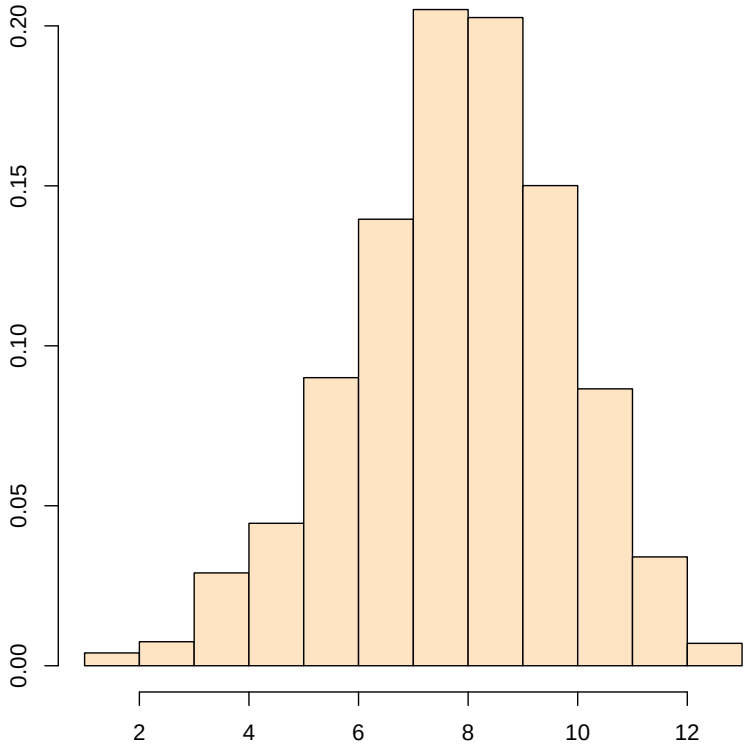


Figure 5.15: Nonprofit organization data: the distribution of the number of clusters.

Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
0	50	4	1	2	5	23	2	1	6.8
15	46	7	9	3	0	41	2	3	12.4
0	67	6	0	0	0	67	11	26	7.5
12	66	2	7	2	4	29	1	12	11.4
0	50	5	5	4	2	40	3	24	15.5
20	86	2	4	0	9	44	1	15	8.8
15	41	5	5	3	2	102	9	14	10.7
0	60	5	4	9	5	22	1	9	13
14	55	3	16	3	2	94	5	7	12.4
0	55	5	3	2	2	62	4	7	14
0	73	2	15	2	2	55	8	5	5.7
9	77	7	7	4	1	67	10	4	5.1
5	63	4	6	31	3	83	25	4	4.2
12	74	4	6	5	1	54	8	8	8.2
21	48	3	4	11	5	21	1	11	15
0	80	4	10	0	2	64	11	5	10.7
0	73	4	14	2	2	23	1	15	10
17	48	7	26	7	0	47	4	13	13.5
10	70	1	8	1	2	80	19	7	4.4
0	71	5	7	0	0	41	3	12	9.1
10	54	5	8	3	2	23	1	6	15
0	54	4	18	18	2	44	1	23	17
0	86	2	7	1	0	64	7	6	5.5
18	72	5	6	7	2	45	2	0	10
8	88	2	10	3	1	65	9	12	6.1
0	45	4	2	12	0	70	9	4	5.6
0	75	4	2	0	0	37	1	6	11.3
0	50	4	10	3	2	63	6	3	14
0	73	2	6	3	2	57	7	7	14
0	74	5	5	2	3	54	4	3	10.6
5	85	1	13	0	1	70	18	10	4.9
12	71	5	3	0	2	63	12	3	9.9
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 5.10: Nonprofit organization data: the parts of the data samples from the original 200 data samples.

Chapter 6

Conclusions and Future Work

Our method gives a new way to estimate the regression function where unknown clusters exist in the data by combining the technique of estimating clusters by Dirichlet mixtures and then estimating regression in each cluster. In simulation experiments in Chapter 5, we can conclude DM method works well compared to other nonparametric methods in most cases. The DM regression method discussed in this thesis exploits the existence of clusters in the data. The DM method may be useful in marketing data or microarray data. We were unable to obtain permission to use actual marketing data due to proprietary issues. However, the program used in this thesis is a free R program which anyone can access. Muller *et al.*(1996) developed a DM based regression estimate by estimating the joint density of $(\boldsymbol{X}, Y)$. But, if we are interested in getting the regression function only, DM method discussed here would be a simpler method. Especially, in higher dimension our method works well no matter how large the dimension is, even in the large p, small n problems.

We have indicated an argument why the proposed method is expected to give correct estimates asymptotically. In the future, we like to construct explicit proof although that seems to be a challenging problem. For the large p, small n problem, we only considered one simple restricted case where all the regressor variables are independent and have the same variance σ^2. More general models such as varying variance or dependence may be considered.

Bibliography

[1] Antoniak, C.E. (1974). Mixtures of Dirichlet Processes With Applications to Nonparametric Problems. *The Annals of Statistics* **2**: 1152–1174.

[2] Barron, A.R., Schervish, M. and Wasserman, L. (1999). The consistency of posterior distributions in non parametric problems. *The Annals of Statistics* **27**: 536–561.

[3] Breiman, L., Friedman, J.H., Olshen, R. and Stone, C.J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth.

[4] Carter, C.K. and Kohn, R. (1997). Semiparametric Bayesian inference for time series with Mixed Spectra. *Journal of the Royal Statistical Society B* **59**: 255–268.

[5] Chambers, J.M., Cleveland, S., Kleiner, B. and Tukey, A.P. (1983). *Graphical Methods for Data Analysis.* Boston: Duxbury.

[6] Choi, T. and Schervish, M. J. (2004). Posterior Consistency in Nonparametric Regression problems under Gaussian process priors. : Technical Report 809, Department of Statistics, Carnegie Mellon University.

[7] Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions:

estimating the correct degree of smoothing by the method of generalized crossvalidation. *Numerical Mathematics* **31**: 377–403.

[8] Cressie, N.A.C (1993). *Statistics for Spatial Data*. Wiley.

[9] Clyde, M. and George, E. (2000). Flexible empirical Bayes estimation for wavelets. *Journal of the Royal Statistical Society B* **62**: 681–698.

[10] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* **39**: 1–38.

[11] Denison, D.G.T., Mallick, B.K. and Smith, A.F.M (1998a). Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society B* **60**: 333–350.

[12] Denison, D.G.T., Mallick, B.K. and Smith, A.F.M (1998b). Bayesian MARS. *Statistical Computing* **8**: 337–346.

[13] DeSarbo, W.S. and Cron, W.L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification* **5**: 249–282.

[14] Diebolt, J. and Robert, C.P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society B* **39**: 363–375.

[15] Donoho, D.L. (1988). One-sided inference about functionals of a density. *The Annals of Statistics* **16**: 1390–1420.

[16] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics* **32**: 407–499.

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

[18] Escobar, M.D. and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association* **90**: 577–588.

[19] Escobar, M.D. and West, M. (1998). Computing Bayesian Nonparametric Hierarchical Models. *Practical Nonparametric and Semiparametric Bayesian Statistics*, lecture Notes in Statistics 133 (Dipak Dey *et al.*, eds,), Springer-Verlag, New-York.

[20] Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems . *The Annals of Statistics* **1**: 209–230.

[21] Fu, W. (1998). Penalized regression: The bridge versus the LASSO. *Journal of Computational Graphical Statististics* **7**: 397–416.

[22] Friedman, J. (1991). Multivariate adaptive regression splines (with discussion). *The Annals of Statistics* **19**: 1–141.

[23] Geweke, J. and Keane, M. (2005). Smoothly Mixing Regressions. Bayesian conference in Econometrics and Statistics. (Available from http://www.olin.wustl.edu /faculty /chib /sbies /uploads /geweke_john.pdf)

[24] Ghosal, S., Ghosh, J.K. and Ramamoorthi, R.V. (1999). Posterior Consistency of Dirichlet Mixtures in Density estimation. *The Annals of Statistics* **27**: 143–158.

[25] Ghosal, S. and van der Vaart, A.W. (2001). Entropies and rates of convergence for Bayes and maximum likelihood estimation for mixture of normal densities.

The Annals of Statistics **29**: 1233–1263.

[26] Hoerl, A. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**: 55–67.

[27] Hoerl, A. and Kennard, R. (1988), Ridge regression, *Encyclopedia of Statistical Sciences*, Vol. 8, Wiley, New York, pp. 129–136.

[28] Liu, J.S. (1996). Nonparametric hierarchical Bayes via sequential imputations. *The Annals of Statistics* **24**: 911–930.

[29] MacKay, D.J.C. (1999). Introduction to Gaussian processes. Technical report. Cambridge University. (Available from http://wol.ra.phy.cam.ac.uk /mackay/gpB.pdf)

[30] Muller, P, Erlkanli, A and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83**: 67–79.

[31] Neal, R.M. (1997). Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. Technical report. University of Toronto. (Available from http://www.cs.toronto.edu/∼radford/)

[32] Omre, H. (1987). Bayesian kriging - merging observations and qualified guess in kriging. *Mathematical Geology* **19**: 25–39.

[33] Priebe, C.E. (1994). Adaptive mixtures. *Journal of the American Statistical Association* **89**: 796–806.

[34] Rand, William M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association* **66**: 846–850.

[35] Rasmussen, C.E. (1996). *Evaluation of Gaussian Process and Other Methods for Non-linear Regression*. PhD Thesis, University of Toronto.

[36] Roeder, K. and Wasserman, L. (1997). Practical density estimation using mixtures of normals. *Journal of the American Statistical Association* **92**: 894–902.

[37] Schwartz, L. (1965). On Bayes procedures. *Z. Wahrsch. Verw. Gebiete* **4**: 10–26.

[38] Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society B* **53**: 683–690.

[39] Spath, H. (1979). Algorithm 39: Clusterwise Linear Regression. *Computing* **22**: 367–373.

[40] Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society B* **58**: 267–288.

[41] Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *The Annals of Statistics* **22**: 1701–1762.

[42] von, Mises, R. (1931). Wahrscheinlich keitsrechnung. Springer, Berlin.

[43] Whittle, P. (1994). Curve and periodogram smoothing. *Journal of the Royal Statistical Society B* **19**: 38–47.

[44] Zou, H. and Hastie, T (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B* **67**: 301–320.