UNIVERSITY OF CALIFORNIA
RIVERSIDE

Bayesian Analysis of MTD/BMTD Models

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

by

Huiming Song

December 2011

Dissertation Committee:

Dr. Keh-shin Lii , Chairperson
Dr. Barry C. Arnold
Dr. Gloria Gonzalez-Rivera

The Dissertation of Huiming Song is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I am grateful to my advisor, Dr. Keh-shin Lii. Without his help and guide, I couldn't get the achievement today. During these past several years, he gave me a lot of help, both in my study and my daily life. All the meetings and discussions with Dr. Lii are of great value for me. I'd like to thank Dr. Arnold who has helped me in my research. I'd like to thank Dr. Gloria Gonzalez-Rivera for her attendance of my oral exam and dissertation defense. Also, I should thank Dr. Jun Li, Dr. Subir Ghosh, and Dr. Xinping Cui. I have learned a lot from their classes. I'd like to make a special thank to Dr. Daniel Jeske, on whose class I began to use statistics in real problems. Lastly, I offer my regards to Paula Lemire and Perla Fabelo. Their work makes me to adapt to the new environment quickly.

To my parents, and my wife Jingxian Huang, for all the support.

ABSTRACT OF THE DISSERTATION

Bayesian Analysis of MTD/BMTD Models

by

Huiming Song

Doctor of Philosophy, Graduate Program in Applied Statistics
University of California, Riverside, December 2011
Dr. Keh-shin Lii , Chairperson

In reality many time series are non-linear and non-Gaussian. They show the characters such as flat stretches, bursts of activity and outliers. Univariate and bivariate mixture transition distribution models were introduced to study these time series data. EM algorithm was used for point estimations of parameters. However as is known, for many mixture models, the likelihoods couldn't be maximized since they will go to infinity. The number of mixtures should be prefixed in this way but in many realities it is unknown.

In our research, Bayesian methods are used to solve these problems. When the posterior is obtained, EM algorithm is used to maximize the posterior. Under some conditions these estimates are proved to be consistent. The second method is using MCMC to sample from the posterior and now the number of mixtures itself can be treated as a random variable. Two methods for MCMC sampling are used. The first is Birth-Death process: if a birth happens, a new mixture component is added; if a death happens, an existing mixture will be removed. The second is Dirichlet process mixtures where we choose Dirichlet process priors for the parameters. When using MCMC, not only point estimations but also interval estimations can be constructed. For all these methods we do simulations to compare Bayesian methods with Non-Bayesian methods and to show the advantages of Bayesian methods.

Contents

Chapter 1

Introduction

1.1 History of Mixture Models

Mixture Transition Distribution (MTD) model was introduced by Le, Martin and Raftery in 1996 to model non-Gaussian time series data with features such as flat stretches, bursts of activity, outliers and change points. MTD can model these behaviors explicitly. This model is simple and easy to study. The MTD model can capture the autoregressive component of the time series, it can also capture other features of the data. For example, occasional outliers may be captured by a mixture component with large variance and small proportion. the flat stretches can be captured by a mixture component with small variance.

Bivariate Mixture Transition Distribution (BMTD) model was introduced by Hassan and Lii in 2006 to model bivariate time series data in general and marked point processes in particular. They extend the MTD model to bivariate model. The BMTD model can not only be used to model time series data but also marked point process data. They introduced a class of bivariate distributions which makes the components of BMTD models having nice marginal densities and conditional densities which are easy to simulate and estimate. The BMTD model can be used in a wide range of applications, such as financial transactions, real-time stock market data, and accidents and events that occur irregularly.

In both just mentioned papers EM algorithm is used for parameter estimation. The mixture model can be treated as missing data problem. Assume each data point comes from one of the k groups (here we think k is fixed and known) with the distribution we selected to model the data. But we don't know which group it comes from. Suppose there is an unobservable variable indicating which group the data is generated. Then it becomes the missing data problem. Our job is to find the value of the parameters (Θ) in the distributions as well as the weight ($\boldsymbol{\pi}$) of each group.

EM algorithm is usually used to estimate parameters for the mixture models. The idea of EM algorithm is easy to understand. It's easy to compute with EM algorithm. Usually AIC or BIC is used to select the correct model and they are easy to calculate. However, it also has some problems: The number of mixtures must be pre-fixed. Most of the time, it finds the local maximum. For many mixture densities, their likelihood will diverge to infinity. Usually BIC is used for exponential family distribution but the mixture model is not in the exponential family in general. For some complicated models, BIC doesn't perform well in model selection.

Let's look at more detailed about the problem of singularity in EM algorithm for mixture models. We want to find the global maximum point of the likelihood function in the EM algorithm. As is known, if we want to use EM algorithm to estimate the parameters, the likelihood cannot be maximized if the parameters in the denominator become zero. For example, the likelihood will be infinity in the normal mixture model, if one of σ_j converges to zero and μ_j comes to one of data point x_t while the corresponding weight π_j is not zero. Under this situation, the likelihood could not be maximized, that is, MLE doesn't exist here. So we need to solve the problem of singularity.

1.1.1 Methods of Parameter Estimation for Mixture Models

We can solve this singularity problem with Bayesian methods. There are two methods can be used here.

Method One and Pros of Bayesian EM

The first is Bayesian EM algorithm. For a fixed k, we choose proper priors for the parameters in the distribution and then get the posterior. Then we will use EM algorithm to maximize the posterior (here the posterior is treated as a function of the parameters and we try to maximize it). Usually the singularity appears because the parameter in the denominator goes to zero in the density. However if we choose a proper prior such as inverse gamma for the parameter σ_j above, the posterior will go to zero rather than infinity when σ_j converges to zero. At last, BIC is used to choose the number of components k. In this way, the priors work like a penalized function for the parameters.

Another pros of this method is that the estimations are consistent. When using EM algorithm without the priors, the MLE does not exist since the likelihood could not be maximized. After using priors to prevent the posterior from being infinity, we can prove the estimations which maximize the posterior are consistent.

Bayesian estimation here is an amend of the non-Bayesian estimation. From theoretical calculation, we can see if we choose non-informative priors for one parameter, then the estimation of this parameter is the same as non-Bayesian method. That is, we can treat the estimations of (Le, Martin, and Raftery, 1996) and (Hassan and Lii, 2006) as special examples in Bayesian frame where uniform priors are used.

Method Two and Pros of MCMC

The alternative method is Markov chain Monte Carlo (MCMC) sampling from the posterior. Birth-Death process method, Dirichlet process method are used here. Using this method, we can both get point estimations as well as interval estimations. Here we are concerned with the analysis of mixture transition models with unknown number of k, and sometimes this k may be of interested itself. Unlike EM algorithm which needs to fix k, MCMC method can simultaneously estimate the parameters in the distribution and the number of components. For this method, after we get the posterior distribution of the parameters, we will try to sample the value of parameters from the posterior.

Usually EM algorithm only gives point estimation of the parameters in the model. However, with MCMC method, after we get the samples of the parameters from the posterior, we can construct confidence intervals for the parameters.

1.2 Summary of My Research

First we introduce the problem of singularity in mixture models in Chapter 2. Since MTD/BMTD models are from mixture models, they also have the problem of singularity and other problems when EM algorithm is used for parameter estimation. In my research, we use Bayesian approach to study MTD and BMTD models. By assigning proper priors to the parameters in the model, the problem of singularities of the posterior can be avoided.

The first method we use is Bayesian approach with EM algorithm. That is, we use EM algorithm to maximize the posterior distribution. Firstly, the number of mixture components k is pre-fixed. Then for this fixed k, EM algorithm is used to maximize the posterior to get the estimations. Then for different k, BIC is used to select the correct model. Because of Bayesian approach, we can solve the problem of singularity and therefore it's possible to maximize the posterior (See Chapter 3 and 4). EM algorithm is easy to understand and requires small amount of computation. Also, we prove that the estimation this way is consistent (Chapter 5).

The second method we use is MCMC. After we get the posterior, we sample from the posterior and then estimate the parameters through the samples. There are two methods we used here to sample.

The first one is called Birth-Death process. It was introduced by Matthew Stephens in 2000. Rather than fixing the value of k, we treat k as a random variable and assign a prior for this k. Birth and Death occur as independent poisson process. When a birth happens, the number of mixture components k will be increased by one.

Conversely, if a death occur, the number of mixture components k will be decreased by one. After we get the samples of k and the parameters, the most appeared number of k (e.g., k_0) will be set as the estimation of true number of mixtures. Then we select all the samples with k equaling to this k_0, and estimate the parameters by the mean of these samples with k equal to k_0. Also, we can construct the interval estimation easily since we have the samples. More detailed information about this method can be found on Chapter 6.

The second mthod is called Dirichlet process mixtures. Sometimes when we get the data, we don't know whether the data is from a mixture model or how many mixtures are there in the model. For each observation data, we assume there is a parameter vector together with that data. Then we group the data by grouping the parameters. The number of unique value of parameters is the estimation of the number of mixtures. One problem is that the parameters are usually continuous, so the probability of two parameter samples equaling to each other is zero. To solve this problem we try to assign a discrete prior for the parameters. We assign a Dirichlet process prior to the parameters. Because of the property of Dirichlet process, the conditional posterior has the clustering feature. It will cluster all of the data through the clustering of parameters. This model gives us the flexibility to choose the random lag order rather than being assigned through the re-parametrization. We discuss the details about this in Chapter 8.

Chapter 2

Singularities of Mixture Models

Finite Mixture Model is widely used in machine learning (Bishop, 2006), cluster analysis (McLachlan, 2000), neural networks (Xu and Jordan, 1996), density estimation and other model constructions. In mixture model context the data are viewed as coming from a mixture of probability distributions, each represents a different cluster. In addition to clustering purposes, finite mixtures of distributions have been applied to a wide variety of statistical problems such as discriminant analysis, image analysis (Jordan, 2006) and survival analysis (Farewell, 1982). To this extent finite mixture models have continued to receive increasing attention from both theoretical and practical points of view.

Because of their flexibility, mixture models are being increasingly exploited as a convenient way to model unknown distribution shapes. For example, Priebe (1994) showed that with $n = 10,000$ observations, a log normal density can be well approximated by a mixture of about 30 normals. A mixture model is able to model quite complex distributions through an appropriate choice of its components to represent accurately the local area of the true distribution. It can thus handle situations where a single parametric family is unable to provide a satisfactory model for local variations in the observed data.

In this section we will study about finite mixture model and its related problems, as well as how to fix these problems. Later we will study infinite mixture models.

2.1 Introduction of Finite Mixture Models

Let $X_1, X_2, \cdots, X_n$ denote a random sample of size n. We suppose the density of X_i can be written in the form

$$f(x_i|\phi) = \sum_{j=1}^{k} \pi_j f_j(x_i|\phi_j) \tag{2.1}$$

where the $f_j(x_i|\phi_j)$ (usually ϕ_j is a vector, here we write it as a scalar) are densities and the π_j are positive quantities that sum to one. The quantities $\pi_1, \pi_2, \cdots, \pi_k$ are called the mixing proportions or weights. As the functions $f_1(x_i|\phi_1),\ \cdots,\ f_k(x_i|\phi_k)$ are densities, it is obvious that (2.1) defines a density. These $f_j(x_i|\phi_j)$ are called the component densities of the mixture. We shall refer to the density (2.1) as a $k-$component finite finite mixture density. In this formulation of the mixture model, the number of components k is considered fixed. But of course in many applications, the value of k is unknown and has to be inferred from the available data, along with the mixing proportions and the parameters in the specified forms of the component densities.

2.1.1 EM Algorithm for Finite Mixture Models

It's common to think mixture models as missing data problems. One way of thinking this is that the data points are coming from the the distribution components we used to model the data. We don't know which component the data comes from. Our objective is to estimate the parameters in each component and the probability of the corresponding component from which the data is generated.

A variety of approaches to the problem of mixture decomposition have been proposed, many of which focus on maximum likelihood methods such as Expectation Maximization (EM) algorithm. Here we briefly consider the EM algorithm.

Assuming $\mathrm{Z}_1, \cdots, \mathrm{Z}_n$ are the hidden random variables where $\mathrm{Z_i} = (Z_{i1}, \cdots, Z_{ik})$ indicates which component that data comes from. For example, $Z_{ip} = 1, Z_{ij} = 0 \ j \neq p$ means data X_i coming from the p_{th} component. We also assume all $\mathbf{Z}$ are independent and are independent of the data $\mathbf{X}$. Now we can write the likelihood function as

$$L = \prod_{i=1}^{n} \prod_{j=1}^{k} (\pi_j f_j(x_i|\phi_j))^{Z_{ij}}. \tag{2.2}$$

EM algorithm is to maximize (2.2) in the following two steps:

1: Expectation Step:

With initial value of the parameters $\phi_j^{(0)}$, $\mathbf{Z}$ is estimated by its conditional expectation as below (McLachlan and Peel, 2000)

$$\tilde{z}_{tj} = \frac{\pi_j f_j \left(x_t|\phi_j^{(0)} \right)}{\sum_{j=1}^{k} \pi_j f_j \left(x_t|\phi_j^{(0)} \right)}, \qquad j = 1 \cdots k.$$

2: Maximization Step:

In this step we need to maximize the likelihood function to get the estimations of the parameters. A common way is to take partial differential of long-likelihood function to parameter ϕ_j like

$$\frac{\partial \ln(L)}{\partial \phi_j} = 0.$$

Solving this equation leads us to the estimation of $\hat{\phi}_j$.

2.1.2 Singularities in Finite Mixture Models

In this section we will discuss the singularity problems of mixture model.

As shown before, we need to maximize the likelihood function to get parameter estimation for mixture models. However, for most of the mixture models, there is the problem called singularity.

Consider the general form of mixture model (2.1)

$$f(x_i|\phi) = \sum_{j=1}^{k} \pi_j f_j(x_i|\phi_j). \tag{2.3}$$

If one of $\pi_j \neq 0$ and the corresponding density function diverges to infinity as the parameter ϕ_j converges to a special point, say $\tilde{\phi}_j$, which is usually at the boundary of the parameter space. That is, if for one of j there is

$$\lim_{\phi_j \to \tilde{\phi}_j} f_j(x_i|\phi_j) = \infty, \tag{2.4}$$

then the density function

$$f(x_i|\phi) = \sum_{j=1}^{k} \pi_j f_j(x_i|\phi_j) \mid_{\ \phi_j \to \tilde{\phi}_j} = \infty. \tag{2.5}$$

That is, the density function will be infinity at some special points.

Let's look at some detailed examples.

1. Finite Normal Mixture Model

   This model is widely used in many areas and can be expressed as

$$\begin{aligned} f(x_i|\mu,\sigma) &= \sum_{j=1}^{k} \pi_j f(x_i|\mu_j,\sigma_j^2) \\ &= \sum_{j=1}^{k} \pi_j \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left\{-\frac{(x_i-\mu_j)^2}{2\sigma_j^2}\right\}. \end{aligned}$$

   If one of $\mu_j = x_i$ and the corresponding $\sigma_j^2 \to 0$ we will have

$$\frac{1}{\sqrt{2\pi}\sigma_j} \exp\left\{-\frac{(x_i-\mu_j)^2}{2\sigma_j^2}\right\}\Bigg|_{\mu_j=x_i,\ \sigma_j^2\to 0} \longrightarrow \infty.$$

   Then the density function

$$f(x_i|\mu,\sigma) = \sum_{j=1}^{k} \pi_j \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left\{-\frac{(x_i-\mu_j)^2}{2\sigma_j^2}\right\}\Bigg|_{\mu_j=x_i,\ \sigma_j^2\to 0} \longrightarrow \infty.$$

2. Finite Log Normal Mixture Model

   The mixture density can be expressed as

$$\begin{aligned} f(x_i|\mu,\sigma) &= \sum_{j=1}^{k} \pi_j f(x_i|\mu_j,\sigma_j^2) \\ &= \sum_{j=1}^{k} \pi_j \frac{1}{\sqrt{2\pi}\sigma_j x_i} \exp\left\{\frac{-(\ln x_i-\mu_j)^2}{2\sigma_j^2}\right\}. \end{aligned}$$

   If one of the $\mu_j = \ln x_i$ and σ_j^2 converges to zero, we will have

$$\frac{1}{\sqrt{2\pi}\sigma_j x_i} \exp\left\{-\frac{(\ln x_i-\mu_j)^2}{2\sigma_j^2}\right\}\Bigg|_{\mu_j=\ln x_i,\ \sigma_j^2\to 0} \longrightarrow \infty.$$

   Then the density function

$$f(x_i|\mu,\sigma) = \sum_{j=1}^{k} \pi_j \frac{1}{\sqrt{2\pi}\sigma_j x_i} \exp\left\{-\frac{(\ln x_i-\mu_j)^2}{2\sigma_j^2}\right\}\Bigg|_{\mu_j=\ln x_i,\ \sigma_j^2\to 0} \longrightarrow \infty.$$

2.2 Simulations

In the above we have discussed the problem of singularity of mixture models of parameter estimation. In the following we will show some simulation examples to demonstrate this.

Suppose the data $X_1,\ X_2,\ \cdots,\ X_n$ are from the following normal mixture model

$$f(x_i|\mu,\sigma) = \sum_{j=1}^{k} \pi_j \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left\{-\frac{(x_i-\mu_j)^2}{2\sigma_j^2}\right\}$$

where π_j is the weight that x_i coming from the j^{th} component, for $j = 1, 2,\ \cdots,\ k$, and $\mu_j,\ \sigma_j^2$ are parameters for the j^{th} component. Previously we have theoretically discussed the problem of singularity in this model. Now we will show when EM algorithm is used for parameter estimation, there does exist the condition that the variance σ_j^2 converges to zero.

It's natural to think the mixture model as the missing data problem. That is, suppose there exist the hidden variable $Z_t = (Z_{t1},\ Z_{t2},\ \cdots,\ Z_{tk})$ which indicates the component that X_t coming from. That is, if $Z_{tj} = 1, Z_{ts} = 0\ s \neq j$ that means the data point X_t comes from the j_{th} component. Now we can write the likelihood function for the data

$$L = \prod_{t=1}^{n}\prod_{j=1}^{k}\left(\pi_j \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left\{-\frac{(x_t-\mu_j)^2}{2\sigma_j^2}\right\}\right)^{Z_{tj}}.$$

EM algorithm has the following two steps.

First is the E step. We need to estimate the missing quantities. Suppose the parameters k, $(\pi_1,\ \pi_2,\ \cdots,\ \pi_k)$, $(\mu_1,\ \mu_2,\ \cdots,\ \mu_k)$ and $(\sigma_1^2,\ \sigma_2^2,\ \cdots,\ \sigma_k^2)$ are known, then the missing quantities Z_t are replaced by its conditional expectations, conditional on the parameters and the observations. The conditional expectation of the j^{th} component of Z_t is just the conditional probability that the observation X_t comes from the j^{th} component of the mixture distribution. Let the conditional expectation of the j^{th} component of Z_t be $\tilde{z}_{tj}$, then we will have

$$\tilde{z}_{tj} = \frac{\pi_j f_j(x_t|\mu_j, \sigma_j^2)}{\sum_{j=1}^{k} \pi_j f_j(x_t|\mu_j, \sigma_j^2)}.$$

Next is the M step. In this step, suppose the estimation of Z_t is given, we need to maximize the likelihood to get estimation of $(\pi_1,\ \pi_2,\ \cdots,\ \pi_k)$, $(\mu_1,\ \mu_2,\ \cdots,\ \mu_k)$ and $(\sigma_1^2,\ \sigma_2^2,\ \cdots,\ \sigma_k^2)$. The simple way is to take partial differential of the log-likelihood function with respect to the parameters

$$\begin{cases} \dfrac{\partial ll}{\partial \pi_j} = 0 \\ \dfrac{\partial ll}{\partial \mu_j} = 0 \\ \dfrac{\partial ll}{\partial \sigma_j^2} = 0. \end{cases} \tag{2.6}$$

where ll is the log-likelihood equals to $\ln(L)$

$$\begin{aligned} ll &= \ln(L) \\ &= \sum_{t=1}^{n}\sum_{j=1}^{k} z_{tj} \ln \pi_j - \sum_{t=1}^{n}\sum_{j=1}^{k} z_{tj} \left(\frac{1}{2} \ln \sigma_j^2 + \frac{(x_t - \mu_j)^2}{2\sigma_j^2} \right). \end{aligned} \tag{2.7}$$

Based on this, we can get the estimators expressed as (proof is at the end of this chapter)

$$\tilde{\pi}_j = \frac{\sum_{t=1}^{n} \tilde{z}_{tj}}{\sum_{t=1}^{n} \sum_{j=1}^{k} \tilde{z}_{tj}},$$

$$\tilde{\mu}_j = \frac{\sum_{t=1}^{n} \tilde{z}_{tj} x_t}{\sum_{t=1}^{n} \tilde{z}_{tj}}, \tag{2.8}$$

$$\tilde{\sigma}_j^2 = \frac{\sum_{t=1}^{n} \tilde{z}_{tj} (x_t - \tilde{\mu}_j)^2}{\sum_{t=1}^{n} \tilde{z}_{tj}}.$$

In the following there are the numerical simulations of the Normal mixture models. We will study the singularities for different kind of conditions and compare them for different sample sizes and different numbers of components. First, consider two normal mixture model, and sample size is $n = 40$. For one component, let its variance be much smaller than the other component. After some iteration, the variance of one component goes to 0 which makes the likelihood approaching infinity. Next we increase the sample size n from 40 to 200 while keeping two components. Now most of the time the algorithm will converge. An average of 15 simulations result is shown below. Finally we increase the mixture components number k from 2 to 5. We can see as the number of components increase, it's more likely there will be singularities. Details are shown below.

2.2.1 Example of Singularity

Consider the normal mixture distribution with 2 components where

$$\boldsymbol{\pi} = (\pi_1, \pi_2) = (.8, .2),$$

$$\boldsymbol{\mu} = (\mu_1, \mu_2) = (1, 9),$$

$$\boldsymbol{\sigma} = (\sigma_1, \sigma_2) = (5, .01),$$

$$f(x_t) = \sum_{j=1}^{2} \pi_j \frac{1}{\sqrt{2\pi}\sigma_j} e^{-(x_t-\mu_j)^2/(2\sigma_j^2)}.$$

First, we will sample 40 samples from the above distribution. For this model, choose the initial values of the parameters as

$$\boldsymbol{\pi}_0 = (\pi_1, \pi_2) = (.7, .3),$$

$$\boldsymbol{\mu}_0 = (\mu_1, \mu_2) = (1.9, 8),$$

$$\boldsymbol{\sigma}_0 = (\sigma_1, \sigma_2) = (4, .1).$$

Using EM algorithm for this data, we can see the variance for the second component goes to zero as iterations go on.

Table 1: Iterations of sigma in the normal distribution, n=40

iteration	σ_1	σ_2
1	5.518	0.191
2	5.475	0.194
$\vdots$	$\vdots$	$\vdots$
50	5.426	0.024
51	5.430	1.15e-29

2.2.2 Increasing Sample Size

Next let the sample size increase from 40 to 200 while keeping other parameters staying the same. The result is

- Repeated the simulation 15 times. Of these 15 simulations, there is once one of the variance converges to zero while other 14 times do not. The following result in Table 2 is from the 14 non-singularity estimation. Take their average as the estimation of the parameters.
- The average result of the 14 simulations are

Table 2: Parameters estimation when n=200

True π	Estimated π	True μ	Estimated μ	True σ	Estimated σ
.800	.952	1.00	2.30	5.00	5.37
.200	.048	9.00	8.01	.01	.02

2.2.3 Increasing the Component Number

Next we let the number of components increase from 2 to 5 and true values of parameters be

$$\boldsymbol{\pi} = (.4,\ .4,\ .1,\ .05,\ .05),$$

$$\boldsymbol{\mu} = (1,\ 9,\ 5,\ 3,\ 12),$$

$$\boldsymbol{\sigma} = (5,\ .01,\ 10,\ .1,\ 1).$$

We use the following initial values for the parameters

$$\boldsymbol{\pi}_0 = (.43,\ .35,\ .12,\ .06,\ .04),$$

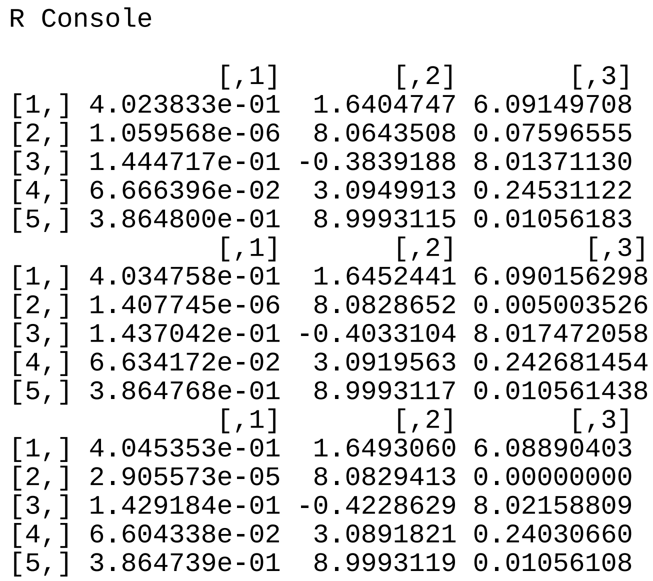
$$\boldsymbol{\mu}_0 = (1.9,\ 8,\ 4.5,\ 4,\ 10),$$

$$\boldsymbol{\sigma}_0 = (4,\ .09,\ 8,\ .16,\ 1.3).$$

Figure 2.1 shows the result of the last three loops of the EM algorithm, from which we can see the variance of the second component goes to zero as the iterations go on. The five rows mean five components of the mixture model. Column 1 stands for the parameter π. column two is the estimation of parameter μ and column three is the standard deviation σ. We can see in column three, for the second row, the standard deviation goes down from 0.0759 to 0.0050 and to 0.0000 at last. That is, the variance of the second component converges to zero. We repeat the simulation 15 times, and all 15 times the singularities happened. That is, if the mixture components increase, it's very likely the singularities occur in the EM algorithm.

2.3 Summary

In this section we reviewed the basics of the mixture model. Also, from theoretical aspect we discussed the problems with mixture models when EM algorithm is used to get the parameter estimation. Singularity is a common problem in the mixture models. Also we did simulations to study the relation between singularity and sample size n as well as mixture component number k. The summary conclusions are given



```
R Console

              [,1]       [,2]       [,3]
[1,] 4.023833e-01  1.6404747 6.09149708
[2,] 1.059568e-06  8.0643508 0.07596555
[3,] 1.444717e-01 -0.3839188 8.01371130
[4,] 6.666396e-02  3.0949913 0.24531122
[5,] 3.864800e-01  8.9993115 0.01056183
              [,1]       [,2]        [,3]
[1,] 4.034758e-01  1.6452441 6.090156298
[2,] 1.407745e-06  8.0828652 0.005003526
[3,] 1.437042e-01 -0.4033104 8.017472058
[4,] 6.634172e-02  3.0919563 0.242681454
[5,] 3.864768e-01  8.9993117 0.010561438
              [,1]       [,2]       [,3]
[1,] 4.045353e-01  1.6493060 6.08890403
[2,] 2.905573e-05  8.0829413 0.00000000
[3,] 1.429184e-01 -0.4228629 8.02158809
[4,] 6.604338e-02  3.0891821 0.24030660
[5,] 3.864739e-01  8.9993119 0.01056108
```

Figure 2.1: Three iteration results for k=5, n=200

below.

- We first let component number be k=2 and sample size be n=40. Let the variance for one component be pretty small which makes this component go to data points simulated from it. Now we see variance for one component goes to zero. That is, *For small sample size, singularities are likely to occur.*

- Secondly, we increase sample size n from 40 to 200. This time we can see that most of the time the EM algorithm will converge and the simulation result is listed on Table 2. That is, as the sample size increases, it's less likely that singularities occur.

- Last, we increase the component number k from 2 to 5, keeping sample size to 200. However, in this case almost every time there is one variance in a component goes to zero thus the likelihood diverges to infinity. Of the 15 replications of the simulation, there are 14 times the result diverges while only once it converges. An example of the divergence is shown in Figure 2.1.

Chapter 3

MTD Normal Model with Bayesian EM Algorithm

3.1 Literature Review

The mixture transition distribution model (MTD) was introduced in 1985 by Raftery for the modeling of high-order Markov chains with a finite state space. Since then it has been generalized and successfully applied to a range of situations, including the analysis of wind directions, DNA sequences and social behavior (Berchtold and Raftery 2002). Mixture Transition Distribution time series model was introduced by LE., Martin and Raftery (1996). It is an extension of general non-Gaussian time series in which the conditional distribution of current observation depending on the previous observations is a mixture of conditional distributions. It can capture non-Gaussian and nonlinear features such as flat stretches , bursts of activity, outliers and change points in a single unified model class. It also performs well in the usual case of Gaussian time series without obvious nonstandard observations. This model is easy to understand, to simulate and it is pretty easy to obtain parameter estimation. When taking into account of prediction, this model is able to obtain a full predictive distribution that can consider the future flat stretches, bursts and outliers.

Since the MTD model is an extension of general mixture model, they use EM algorithm to obtain parameter estimation for the mixture model. When using EM algorithm, firstly the number of mixture components is assumed to be known. That is, the mixture number k is assumed to be known. Then for each observation Y_i it is assumed that there is a hidden vector Z_i which indicates the mixture component that Y_i is coming from. With these assumptions, they can write out the conditional likelihood function L. Then EM algorithm can be applied to maximize the conditional likelihood function to get the parameter estimation. For different k, different models will be constructed. Bayesian Information Criteria (BIC) is used to determine the number of mixture components k.

There are other papers about the MTD models, including extending the univariate normal mixtures to multivariate mixture models (Baudry, Raftery and Celeux 2000). Most of these papers use EM algorithm to estimate the parameters and use BIC to determine the number of mixture components.

However, as is known in normal mixture models, there is a problem in maximizing the likelihood function of mixture models because of singularity. Another problem with EM algorithm is that it needs to predetermine the number of mixture components and then use BIC to determine the number of mixtures. However, most of time, we have no idea how many components are there for the data or maybe even there is no actual mixtures, so usually it's not easy to predetermine what k should be used. The third problem is that BIC is usually used for exponential family model selection. But the mixture model does not belong to exponential family in general. Some other people, such as Wong and Li (2000) suggest other criteria called BIC* for their model. So, most of time there is no universal criteria for model selection here.

Ridolfi and Idier (2000) used penalized maximum likelihood estimation for normal mixture distributions to solve the problem of singularity. In this paper, they studied normal mixture model and used penalized function to avoid the singularities. Since the likelihood function is not bounded because of singularities at the boundary of the parameter domain, MLE for the mixture model has a problem of singularities. They assigned penalized functions for the possible singular parameters. In their paper for normal mixtures, penalized function is used for the parameter σ^2. By using EM algorithm, the M-Step is to maximize the likelihood function × penalized function. In this way, singularities can be avoided duo to the penalized function. In this example, they compared the penalized function method with Hathaway's (1985) constrained method (that is, make restriction with the parameter $\sigma_1, \ \ \sigma_2$, i.e. $\sigma_1/\sigma_2 \geq c > 0$).

In this section, we will present the Bayesian method for the MTD models. Under the Bayesian framework, we can avoid the singularity. The penalized function method can be treated as a special case of Bayesian method where the priors are non-informative for parameters unrelated with singularities. In this section we first choose proper priors for the parameters in the MTD model. After we get the posterior we use EM algorithm to maximize the corresponding posterior distribution. At last we use BIC to select k. This method solves the problem of singularity but it still has the shortcomings of EM algorithm, such as converging to a local maximum, or the problem that BIC is not suitable. We can show mathematically there is no singularity problem using EM algorithm here.

3.2 Introduction of MTD Model and Problems in EM Algorithm

3.2.1 Introduction of MTD Model

Mixture Transition Distribution (MTD) time series model was introduced by LE., Martin and Raftery (1996). Here we will study the mixture transition distribution of normal models. Suppose we have data $x_1, x_2, \cdots, x_n$, the conditional distribution of x_t current given the previous information $x^{t-1} = (x_{t-1}, x_{t-2}, \cdots, x_1)$, can be expressed as

$$\begin{aligned} f(x_t|x^{t-1}, \phi) &= \sum_{j=1}^{k} \pi_j f_j(x_t|x^{t-1}, \theta_j, \sigma_j^2) \\ &= \sum_{j=1}^{k} \frac{\pi_j}{\sqrt{2\pi}\sigma_j} \exp\left\{ -\frac{(x_t - \theta_j x_{t-j})^2}{2\sigma_j^2} \right\}, \quad t = k+1, \cdots, n. \qquad (3.1) \end{aligned}$$

$\phi = (\phi_1, \ \cdots, \ \phi_k)$ $(\phi_j = \{\pi_j, \theta_j, \sigma_j^2\})$ are the parameters in the MTD density.

This is a mixture distribution of k components. For the j^{th} component, the density depends of the j^{th} lag information with mean μ_j equaling to $\theta_j x_{t-j}$ and variance

σ_j^2. The parameter space is denoted as Φ. We have

$$\Phi = \{\phi_j = \{\pi_j, \theta_j, \sigma_j^2\} |\ \pi_j \in \mathbb{R}_+,\ \sum_{j=1}^{k} \pi_j = 1,$$
$$\theta_j \in \mathbb{R};\ \sigma_j^2 \in \mathbb{R}_+/\{0\}, \text{for } j = 1,\ 2, \cdots, k\}.$$

Given the data $\mathbf{X}$, the maximum likelihood estimation of the mixture parameters is defined as

$$\hat{\phi}_T | f(\mathbf{x}, \hat{\phi}_T) = \sup_{\phi \in \Phi} f(\mathbf{x}, \phi)$$

where $f(\mathbf{x}, \phi)$ is the likelihood function

$$f(\mathbf{x}, \phi) = \prod_{t=k+1}^{n} f(x_t | x^{t-1}, \phi) = \prod_{t=k+1}^{n} \sum_{j=1}^{k} \pi_j f_j(x_t | x^{t-1}, \theta_j, \sigma_j^2). \tag{3.2}$$

3.2.2 Degeneracy of Likelihood Function

For most of the mixture models, there is a well-known problem of likelihood function degeneracy. Let us consider an easy condition of the previous model (3.1): suppose there are only two mixtures, then we will rewrite the likelihood function as

$$f(\mathbf{x}, \phi) = \prod_{t=3}^{n} \left(\frac{\pi_1}{\sqrt{2\pi}\sigma_1} \exp\left\{ -\frac{(x_t - \theta_1 x_{t-1})^2}{2\sigma_1^2} \right\} \right.$$
$$\left. + \frac{\pi_2}{\sqrt{2\pi}\sigma_2} \exp\left\{ -\frac{(x_t - \theta_2 x_{t-2})^2}{2\sigma_2^2} \right\} \right). \tag{3.3}$$

By intuition, we can find the degeneracy is due to the variance parameter converging to zero in the denominator. In fact, some estimators such as $\{\pi_2 \neq 0,\ \theta_2 = x_t/x_{t-2}$ and $\sigma_2^2 = 0\}$ will yield singularities in the sense that the likelihood function f goes to infinity. In fact, when we consider the parameters values on the boundary of the parameter space Φ, denoted as $\partial\Phi$, we can get the singularities if the parameters ϕ approaching one of the corresponding values on the boundary of Φ.

Property 1. *Let us consider the likelihood function (3.2), then*

$$\forall x \in \mathbb{R}, \exists \phi^0 \in \partial\Phi, \text{ such that } \lim_{\phi \to \phi^0} f(\mathbf{x}, \phi) = +\infty$$

where Φ is the parameter space, $\partial\Phi$ is the boundary of parameter space, $\phi^0 = \{\pi_j \neq 0;\ \theta_j = x_t/x_{t-j};\ \sigma_j^2 = 0\} \in \partial\Phi$ is a point on the boundary of the parameter space.

3.2.3 Bayesian Method and Its Pros

From above we have shown if we want to maximize the likelihood function, we will have to face the problem of singularity and therefore it's difficult to get MLE for the MTD model.

Now we will go to Bayesian framework and we will show under Bayesian framework, we can avoid the problem of singularity.

For the MTD model (3.1), because the degeneracy comes from σ_j^2 on the denominator, we will choose proper conjugate priors for the variance σ_j^2. With this proper conjugate priors, not only the posterior becomes easy to calculate, but also the posterior will never go to infinity.

Here we choose Inverse Gamma $IG(\alpha_j, \beta_j)$ as the conjugate priors for σ_j^2

$$IG(\sigma_j^2 | \alpha_j, \beta_j) \propto \left(\frac{1}{\sigma_j^2}\right)^{\alpha_j} \exp\left\{-\frac{\beta_j}{\sigma_j^2}\right\}. \tag{3.4}$$

So, now the posterior becomes

$$\begin{aligned}
posterior &= likelihood \times priors \\
&= f(\mathbf{x}, \phi) \times \prod_{j=1}^{k} IG(\sigma_j^2|\alpha_j, \beta_j) \\
&= \prod_{t=k+1}^{n} f(x_t|x^{t-1}, \phi) IG(\sigma_j^2|\alpha_j, \beta_j) \\
&= \prod_{t=k+1}^{n} \sum_{j=1}^{k} \pi_j f_j(x_t|x^{t-1}, \theta_j, \sigma_j^2) \times \prod_{j=1}^{k} \frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)} \left(\frac{1}{\sigma_j^2}\right)^{\alpha_j} \exp\left\{-\frac{\beta_j}{\sigma_j^2}\right\} \\
&= \prod_{t=k+1}^{n} \sum_{j=1}^{k} \frac{\pi_j}{\sqrt{2\pi}\sigma_j} \exp\left\{-\frac{(x_t - \theta_j x_{t-j})^2}{2\sigma_j^2}\right\} \\
&\quad \times \prod_{j=1}^{k} \frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)} \left(\frac{1}{\sigma_j^2}\right)^{\alpha_j} \exp\left\{-\frac{\beta_j}{\sigma_j^2}\right\}. \qquad (3.5)
\end{aligned}$$

Now we can see if one of σ_j^2 goes to zero, although the likelihood part will go to infinity with linear speed, the prior part will go to zero with exponential speed. Combined together, we will have the following conclusion.

Property 2. *In maximizing the posterior distribution (3.5), then even if the parameter ϕ is on the boundary, such as σ_j approaching zero, the posterior will not approach infinity.*

Proof: See appendix of proof of this chapter.

From Property 2, we know that the posterior will all the time be finite. This gives us the possibility to maximize the posterior distribution to get the parameter estimation. In the next section we will use EM algorithm to maximize the posterior and use BIC for model selection.

3.3 Bayesian Method with EM Algorithm for MTD Models

In this section we will restudy the MTD model with Bayesian method. Under Bayesian framework, we first assign priors for the parameters in the MTD models. In order to calculate the posterior easily and conveniently, we will assign conjugate priors for the parameters. Combined with the observational data, we can get the posterior and then EM algorithm is used to get parameter estimation by maximizing the posterior distribution.

Assuming $\mathrm{Z}_1, \cdots, \mathrm{Z}_n$ are the hidden random variables where $\mathrm{Z_i} = (Z_{i1}, \cdots, Z_{ik})$ indicates which the component that data comes from. For example, $Z_{ip} = 1, Z_{ij} = 0\ j \neq p$ means data X_i coming from the p^{th} component. We also assume all $\mathrm{Z}_i's$ are independent and are independent of the data X. Now we can write the (conditional) likelihood function as

$$L = \prod_{t=k+1}^{n} \prod_{j=1}^{k} \left(\frac{\pi_j}{\sqrt{2\pi}\sigma_j} \exp\left\{ -\frac{(x_t - \theta_j x_{t-j})^2}{2\sigma_j^2} \right\} \right)^{z_{tj}}. \tag{3.6}$$

The prior for $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ is chosen as non-informative prior. The prior for σ_j^2 is chosen as (3.4). Based on this, the posterior distribution becomes

$$P \propto \prod_{t=k+1}^{n} \prod_{j=1}^{k} \left(\frac{\pi_j}{\sqrt{2\pi}\sigma_j} \exp\left\{ -\frac{(x_t - \theta_j x_{t-j})^2}{2\sigma_j^2} \right\} \right)^{z_{tj}} \times \prod_{j=1}^{k} \frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)} \left(\frac{1}{\sigma_j^2} \right)^{\alpha_j} \exp\left\{ -\frac{\beta_j}{\sigma_j^2} \right\}.$$

So, the logarithm of the posterior distribution is

$$\begin{aligned} ll &= \log(posterior) \\ &= \sum_{t=k+1}^{n}\sum_{j=1}^{k} z_{tj}\log\pi_j + \sum_{t=k+1}^{n}\sum_{j=1}^{k} z_{tj}\log f_j(x_t|x^{(t-1)}, \theta_j, \sigma_j^2) \\ &\quad + \sum_{j=1}^{k}\log IG(\sigma_j^2|\alpha_j, \beta_j) \\ &= \sum_{t=k+1}^{n}\sum_{j=1}^{k} z_{tj}\log\pi_j + \sum_{t=k+1}^{n}\sum_{j=1}^{k} z_{tj}\left(-\log\sqrt{2\pi}\sigma_j - \frac{(x_t-\theta_j x_{t-j})^2}{2\sigma_j^2}\right) \\ &\quad + \sum_{j=1}^{k}\left(\alpha_j\log\beta_j - \log(\Gamma(\alpha_j)) - \alpha_j\log\sigma_j^2 - \frac{\beta_j}{\sigma_j^2}\right). \end{aligned} \tag{3.7}$$

Next we will use EM algorithm to get parameter estimation.

First is the E step we need to estimate the missing quantities. Suppose the parameters $(\theta_1, \cdots, \theta_k)$ and $(\sigma_1^2, \cdots, \sigma_k^2)$ are known, then the missing quantities of Z_t are replaced by their conditional expectations, conditional on the parameters and the observations. The conditional expectation of the j^{th} component of Z_t is just the conditional probability that the observation X_t comes from the j^{th} component of the mixture distribution, conditional on the parameters and observations. Let the conditional expectation of the j^{th} component of Z_t be $\tilde{z}_{tj}$, then we will have

$$\tilde{z}_{tj} = \frac{\pi_j f_j(x_t|x^{(t-1)})}{\sum_{j=1}^{k}\pi_j f_j(x_t|x^{(t-1)})} \tag{3.8}$$

for $j = 1,\ 2, \cdots, k$.

Next is the M step. Now suppose we know the value of Z_t, what we need to do is to maximize the posterior distribution to get the estimation of the parameters of $(\theta_1, \cdots, \theta_k)$ and $(\sigma_1^2, \cdots, \sigma_k^2)$. To maximize it, we know the estimations of parameters should satisfy these equations

$$\begin{cases} \dfrac{\partial ll}{\partial \pi_j} = 0, \\ \dfrac{\partial ll}{\partial \theta_j} = 0, \\ \dfrac{\partial ll}{\partial \sigma_j^2} = 0. \end{cases} \tag{3.9}$$

Take partial difference to $\pi_j,\ \theta_j$ and σ_j^2, for $j = 1,\ 2, \cdots, k$, we will get

$$\frac{\partial ll}{\partial \pi_j} = 0 \Longrightarrow \hat{\pi}_j = \frac{\sum_{t=k+1}^{n} \tilde{z}_{tj}}{\sum_{t=k+1}^{n} \sum_{j=1^k} \tilde{z}_{tj}}, \tag{3.10}$$

$$\frac{\partial ll}{\partial \theta_j} = 0 \Longrightarrow \hat{\theta}_j = \frac{\sum_{t=1}^{n} \tilde{z}_{tj} x_t x_{t-j}}{\sum_{t=1}^{n} \tilde{z}_{tj} x_{t-j}^2}. \tag{3.11}$$

This Bayesian estimator of $\hat{\theta}_j$ is the same as non-Bayesian method estimation of $\hat{\theta}_j$.

$$\begin{aligned} \frac{\partial ll}{\partial \sigma_j^2} = 0 \Longrightarrow \hat{\sigma}_j^2 &= \frac{\sum_{t=1}^{n} \tilde{z}_{tj}(x_t - \hat{\theta}_j x_{t-j})^2 + 2\beta_j}{\sum_{t=1}^{n} \tilde{z}_{tj} + 2\alpha_j} \\ &= \frac{2\beta_j}{\sum_{t=1}^{n} \tilde{z}_{tj} + 2\alpha_j} + \frac{\sum_{t=1}^{n} Z_{tj}(x_t - \hat{\theta}_j x_{t-j})^2}{\sum_{t=1}^{n} \tilde{z}_{tj} + 2\alpha_j}. \end{aligned} \tag{3.12}$$

If we don't use Bayesian method, the estimator is

$$\hat{\sigma}_j^2 = \frac{\sum_{t=1}^{n} \tilde{z}_{tj}(x_t - \theta_j x_{t-j})^2}{\sum_{t=1}^{n} \tilde{z}_{tj}}. \tag{3.13}$$

Calculation of (3.10), (3.11) and (3.12) will be shown on last section of this Chapter.

If we pay more attention to (3.12) and (3.13), we see that in (3.12), as the sample size n increase, $\sum_{t=1}^{n} \tilde{z}_{tj}$ will also increase consequently. If we get a large sample size n, the $\sum_{t=1}^{n} \tilde{z}_{tj} + 2\alpha_j$ will be close to $\sum_{t=1}^{n} \tilde{z}_{tj}$. $\frac{2\beta_j}{\sum_{t=1}^{n} \tilde{z}_{tj}+2\alpha_j}$ will be close to zero since $\sum_{t=1}^{n} \tilde{z}_{tj}$ in the denominator is very large compared to the fixed β_j in the numerator. And $\frac{\sum_{t=1}^{n} Z_{tj}(x_t-\theta_j x_{t-j})^2}{\sum_{t=1}^{n} \tilde{z}_{tj}+2\alpha_j}$ will go to $\frac{\sum_{t=1}^{n} Z_{tj}(x_t-\theta_j x_{t-j})^2}{\sum_{t=1}^{n} \tilde{z}_{tj}}$. That is, (3.12) will approach to (3.13) as sample size n increase. That is, if we increase the sample size, there will be of little difference between Bayesian and non-Bayesian method in the estimation when there is no singularity appears. However, as is shown, Bayesian estimator will always exclude the condition of variances approaching zero while non-Bayesian method cannot guarantee this. This is the superiority of Bayesian method.

3.4 Simulation Result for MTD Normal Model

3.4.1 Simulation Results

In the simulation, we set the true number of mixture components $k = 3$. Sample size is $n = 200$. The simulation is repeated for 100 times. The true model is

$$\begin{aligned} f(x_t, \phi) = & \frac{.1}{\sqrt{2\pi} \times .1} \exp\left\{-\frac{(x_t - .2x_{t-1})^2}{2 \times .1^2}\right\} \\ & + \frac{.7}{\sqrt{2\pi} \times 1} \exp\left\{-\frac{(x_t - .3x_{t-2})^2}{2 \times 1^2}\right\} \\ & + \frac{.2}{\sqrt{2\pi} \times 5} \exp\left\{-\frac{(x_t + 2.5x_{t-3})^2}{2 \times 5^2}\right\}. \end{aligned}$$

The simulation result is given as below.

k=3 n=200 Non-Bayesian Method

Of all 100 simulations, there is 5 times singularity will happen and other 95 times there is no singularity. The simulation result is shown below (in the brackets is the true value of the model).

Estimation of the parameters

k	π	θ	σ
1	0.133031 (.1)	0.166431 (.2)	0.120865 (.1)
2	0.671288 (.7)	0.281730 (.3)	0.966687 (1)
3	0.195679 (.2)	-2.311235 (-2.5)	5.064729 (5)

standard errors of the Estimation

k	π	θ	σ
1	0.056418	0.100086	0.105686
2	0.165906	0.069373	0.248471
3	0.060096	0.686751	2.019126

k=3 n=200 Bayesian Method

Next are the simulation of the same true model and the same initial value of the parameters as the simulation above. Using Bayesian method, we can avoid the appearance of singularity in the simulation. Also, we see here the Bayesian method estimation is better than non-Bayesian method. The simulation result is:

Estimation of the parameters

k	π	θ	σ
1	0.107921 (.1)	0.194684 (.2)	0.113812 (.1)
2	0.693693 (.7)	0.303651 (.3)	1.010320 (1)
3	0.200126 (.2)	-2.440371 (-2.5)	5.031268 (5)

standard errors of the Estimation

k	π	θ	σ
1	0.053266	0.089513	0.050552
2	0.052318	0.026148	0.083693
3	0.032019	0.406630	0.775302

From the simulation we can see two advantages of Bayesian method. First, there is no singularity in Bayesian method. Second, estimations from Bayesian method is better than non-Bayesian method; standard errors of estimations from Bayesian method is smaller then non-Bayesian method.

3.4.2 Comparison Study

In this section we will restudy the example Le, Martin and Raftery used in their paper (1996). It's a normal MTD model with 3 mixture components. We will compare their result with the result of Bayesian method here. Also, for the EM algorithm, we will compare the result using good initial values (close to our true value) with the result using bad initial values (values far from the true value and the initial value of standard errors is close to zero). True model is

$$\begin{aligned} f(x_t, \phi) = & \frac{.4}{\sqrt{2\pi} \times 1} \exp\left\{-\frac{(x_t - .3x_{t-1})^2}{2 \times 1^2}\right\} \\ & + \frac{.4}{\sqrt{2\pi} \times 1} \exp\left\{-\frac{(x_t - .3x_{t-2})^2}{2 \times 1^2}\right\} \\ & + \frac{.2}{\sqrt{2\pi} \times 5} \exp\left\{-\frac{(x_t + 2.5x_{t-3})^2}{2 \times 5^2}\right\}. \end{aligned}$$

Good Initial Value Simulation Comparison

First we set the initial values with all σ_j far away from zero. Specifically they are chosen as

Good Initial Value

k	π	θ	σ
1	.32	.1	.5
2	.50	.5	.5
3	.18	-2.0	3.5

The simulation result is list below.

Non-Bayesian with Good initial value

Of 100 replications, there is one time the variance goes to zero. Other 99 times singularity doesn't appear. This result is close to the result as they got in their paper.

Estimation of the parameters

k	π	θ	σ
1	0.3936652	0.2850216	0.9603959
2	0.3946651	0.3028768	0.9377454
3	0.1916697	-2.3269386	4.7824925

standard errors of the Estimation

k	π	θ	σ
1	0.11204791	0.07550024	0.2448749
2	0.11196956	0.07025475	0.2391312
3	0.05069433	0.64051946	0.9448469

Bayesian with Good initial value

Of 100 replications, there is no problem of singularity.

Estimation of the parameters

k	π	θ	σ
1	0.3947873	0.2871532	0.9256657
2	0.4020719	0.3107894	0.9375738
3	0.2031407	-2.4941271	4.8240909

standard errors of the Estimation

k	π	θ	σ
1	0.11007369	0.08452048	0.1983980
2	0.10772111	0.05979850	0.1976612
3	0.03703981	0.39752495	0.7395327

We can see here for Bayesian method with good initial values, of all 9 estimations, 7 estimations with Bayesian method are better then estimators with non-Bayesian method. For the standard errors, there are 8 of with in Bayesian method smaller then non-Bayesian method.

Bad Initial Value Simulation Compare

Next we use the initial value with some for σ_j close to zero. The initial values are given below.

Bad Initial Value

k	π	θ	σ
1	.32	.1	.05
2	.50	.5	.5
3	.18	-2.0	3.5

Non-Bayesian with Bad initial value

Of 100 replications, there are 5 times the variance will go to zero. Because the initial value is poor, the simulation is far away from the true value.

Estimation of the parameters

k	π	θ	σ
1	0.1705309	0.1663157	0.3764691
2	0.5781841	0.2465586	1.1286935
3	0.2012851	-2.2513687	5.0595137

standard errors of the Estimation

k	π	θ	σ
1	0.2019159	0.10032554	0.4332319
2	0.2339350	0.08407647	0.5985411
3	0.0648519	0.63069761	2.0129913

Bayesian with Bad initial value

Of 100 replications, no time the variance will go to zero.

Estimation of the parameters

k	π	θ	σ
1	0.2286117	0.1928289	0.5429712
2	0.5517805	0.2617159	1.1445907
3	0.2196078	-2.4058028	5.1568373

standard errors of the Estimation

k	π	θ	σ
1	0.16505776	0.11691836	0.2835764
2	0.16518295	0.07399314	0.2279577
3	0.04167456	0.30918145	1.7339096

From this simulation, it shows if the simulation starts with poor initial valurs, estimators from both Bayesian and non-Bayesian method are poor. This suggest us to be careful when choosing initial values for EM algorithm. If we begin from some bad initial values, the estimations might be poor. However it seems that the Bayesian method gives better estimates than those of non-Bayesian method.

3.5 Summary

In this chapter, we use Bayesian method for MTD models. For fixed k (number of components), we choose proper priors for parameters, then use EM to maximize the posterior to obtain parameter estimation.

The advantages of using Bayesian EM algorithm are:

- It can eliminate the problem of singularities.
- The estimation is consistent under some restrictions for the parameters' prior (we will prove this in Chapter 5).
- With proper prior, Bayesian estimation performs better than non-Bayesian method if the sample size is not too small.
- As n (sample size) increases, Bayesian method is close to non-Bayesian method but without singularity problem.

3.6 Proofs

3.6.1 Proof of Property (2)

From (3.5) we know the posterior

$$
\begin{aligned}
&= \prod_{t=k+1}^{n} \sum_{j=1}^{k} \frac{\pi_j}{\sqrt{2\pi}\sigma_j} \exp\left\{-\frac{(x_t-\theta_j x_{t-j})^2}{2\sigma_j^2}\right\} \\
&\quad \times \prod_{j=1}^{k} \frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)} \left(\frac{1}{\sigma_j^2}\right)^{\alpha_j} \exp\left\{-\frac{\beta_j}{\sigma_j^2}\right\} \\
&\leq \prod_{t=k+1}^{n} \left(\sum_{j=1}^{k} \frac{\pi_j}{\sqrt{2\pi}\sigma_j}\right) \times \prod_{j=1}^{k} \frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)} \left(\frac{1}{\sigma_j^2}\right)^{\alpha_j} \exp\left\{-\frac{\beta_j}{\sigma_j^2}\right\} \\
&= \prod_{t=k+1}^{n} \left(\sum_{j=1}^{k} \left(\frac{1}{\sigma_j^2}\right)^{1/2}\right) \times \prod_{j=1}^{k} \left(\frac{1}{\sigma_j^2}\right)^{\alpha_j} \exp\left\{-\frac{\beta_j}{\sigma_j^2}\right\} \times \prod_{j=1}^{k} \frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)} \\
&= d_0 \times \prod_{t=k+1}^{n} \left(\sum_{j=1}^{k} \left(\frac{1}{\sigma_j^2}\right)^{1/2}\right) \times \prod_{j=1}^{k} \left(\frac{1}{\sigma_j^2}\right)^{\alpha_j} \exp\left\{-\frac{\beta_j}{\sigma_j^2}\right\} \\
&= d_0 \times \prod_{t=k+1}^{n} \left(\sum_{j=1}^{k} \left(\frac{1}{\sigma_j^2}\right)^{1/2} \times \prod_{j=1}^{k} \left(\frac{1}{\sigma_j^2}\right)^{\alpha_j \frac{k+1}{nk}} \exp\left\{-\frac{\beta_j}{\sigma_j^2}\frac{k+1}{nk}\right\}\right) \\
&= d_0 \times \prod_{t=k+1}^{n} \Bigg(\sum_{j=1}^{k} \Bigg(\frac{1}{\sigma_j^2} \times \prod_{j=1}^{k} \left(\frac{1}{\sigma_j^2}\right)^{2\alpha_j \frac{k+1}{nk}} \exp\left\{-\frac{2\beta_j}{\sigma_j^2}\frac{k+1}{nk}\right\}\Bigg)\Bigg)^{1/2}\Bigg) \\
&= d_0 \times \prod_{t=k+1}^{n} \Bigg(\sum_{j=1}^{k} \Bigg(\prod_{j=1}^{k} \left(\frac{1}{\sigma_j^2}\right)^{2\alpha_j \frac{k+1}{nk} + \frac{1}{k}} \exp\left\{-\frac{2\beta_j}{\sigma_j^2}\frac{k+1}{nk}\right\}\Bigg)\Bigg)^{1/2}\Bigg) \quad .
\end{aligned} \tag{3.14}
$$

The first $\leq$ sign in the above is duo to $\exp\left\{-\frac{(x_t-\theta_j x_{t-j})^2}{2\sigma_j^2}\right\} \leq 1$. And we denote $d_0 = \prod_{j=1}^{k} \frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)}$. Because

$$
\left(\frac{1}{\sigma_j^2}\right)^{2\alpha_j \frac{k+1}{nk} + \frac{1}{k}} \exp\left\{-\frac{2\beta_j(k+1)}{nk}\frac{1}{\sigma_j^2}\right\}
$$

is finite for any σ_j^2 even if σ_j^2 is approaching its boundary at zero. So, for all σ_j^2, (3.14) is boundary. That is, when using Bayesian method, the posterior will be boundary for any σ_j^2.

3.6.2 Calculations for (3.10), (3.11), (3.12)

From (3.7) we know the log-posterior is

$$\begin{aligned}
ll =& \sum_{t=k+1}^{n}\sum_{j=1}^{k} z_{tj}\log\pi_j + \sum_{t=k+1}^{n}\sum_{j=1}^{k} z_{tj}\left(-\log(\sqrt{2\pi}\sigma_j) - \frac{(x_t-\theta_j x_{t-j})^2}{2\sigma_j^2}\right)\\
&\sum_{j=1}^{k}\left(\alpha_j\log\beta_j - \log(\Gamma(\alpha_j)) - \alpha_j\log\sigma_j^2 - \frac{\beta_j}{\sigma_j^2}\right)\\
=& \sum_{t=k+1}^{n}\sum_{j=1}^{k} z_{tj}\log\pi_j - \sum_{t=k+1}^{n}\sum_{j=1}^{k} z_{tj}\left(\frac{(x_t-\theta_j x_{t-j})^2}{2\sigma_j^2}\right)\\
&-\frac{1}{2}\sum_{t=k+1}^{n}\sum_{j=1}^{k} z_{tj}\log(\sigma_j^2) - \sum_{j=1}^{k}\left(\alpha_j\log\sigma_j^2\right) - \sum_{j=1}^{k}\left(\frac{\beta_j}{\sigma_j^2}\right)\\
&-\sum_{j=1}^{k}\log\sqrt{2\pi} + \sum_{j=1}^{k}\alpha_j\log\beta_j - \sum_{j=1}^{k}\log(\Gamma(\alpha_j)).
\end{aligned}$$

Take partial difference of ll with respect to $\pi_j,\ \theta_j,\ \sigma_j^2$, by setting $\frac{\partial ll}{\partial\pi_j}=0, j=1,\ 2,\ \cdots,\ k-1$, we obtain

$$\sum_{t=k+1}^{n} z_{tj}/\pi_j - \sum_{t=k+1}^{n} z_{tk}/(1-\pi_1-\cdots-\pi_{k-1}) = 0,$$

$$\sum_{t=k+1}^{n} z_{tj}/\pi_j - \sum_{t=k+1}^{n} z_{tk}/\pi_k = 0,$$

$$\sum_{t=k+1}^{n} z_{tj}\pi_k = \sum_{t=k+1}^{n} z_{tk}\pi_j,$$

$$\frac{\pi_j}{\pi_k} = \frac{\sum_{t=k+1}^{n} z_{tj}}{\sum_{t=k+1}^{n} z_{tk}} \quad ,\text{for } j = 1,\ 2,\ \cdots,\ k-1,$$

$$\hat{\pi}_j = \frac{\sum_{t=k+1}^{n} z_{tj}}{\sum_{t=k+1}^{n}\sum_{j=1^k} z_{tj}}. \tag{3.15}$$

By setting $\frac{\partial ll}{\partial \theta_j} = 0$, we obtain

$$\sum_{t=k+1}^{n} z_{tj}\frac{1}{\sigma_j^2}(x_t - \theta_j x_{t-j})x_{t-j} = 0,$$

$$\sum_{t=k+1}^{n} z_{tj}x_t x_{t-j} - \theta_j \sum_{t=k+1}^{n} z_{tj}(x_{t-j})^2 = 0,$$

$$\hat{\theta}_j = \frac{\sum_{t=1}^{n} z_{tj}x_t x_{t-j}}{\sum_{t=1}^{n} z_{tj}x_{t-j}^2}. \tag{3.16}$$

By setting $\frac{\partial ll}{\partial \sigma_j^2} = 0$, we obtain

$$\sum_{t=k+1}^{n} z_{tj}\left(\frac{(x_t-\theta_j x_{t-j})^2}{2(\sigma_j^2)^2}\right) - \frac{1}{2}\sum_{t=k+1}^{n} z_{tj}\frac{1}{\sigma_j^2} + \frac{\alpha_j}{\sigma_j^2} + \frac{\beta_j}{(\sigma_j^2)^2} = 0,$$

$$\left(\sum_{t=k+1}^{n} z_{tj}(x_t-\theta_j x_{t-j})^2 + 2\beta_j\right) - \left(\sum_{t=k+1}^{n} z_{tj} + 2\alpha_j\right)\sigma_j^2 = 0,$$

$$\hat{\sigma}_j^2 = \frac{\sum_{t=1}^{n} Z_{tj}(x_t-\theta_j x_{t-j})^2 + 2\beta_j}{\sum_{t=1}^{n} z_{tj} + 2\alpha_j}$$

$$= \frac{2\beta_j}{\sum_{t=1}^{n} z_{tj} + 2\alpha_j} + \frac{\sum_{t=1}^{n} z_{tj}(x_t-\theta_j x_{t-j})^2}{\sum_{t=1}^{n} z_{tj} + 2\alpha_j}.$$

Chapter 4

BMTD Model with Bayesian EM Algorithm

4.1 Introduction of BMTD Model

Hassan and Lii (2006) introduced Bivariate Mixture Transition Distribution (BMTD) model. It can be used in many different areas such as time series and marked point processes. For marked point processes, usually people will assume the interval times and marks mutually independently distributed. There is no such restriction in the BMTD model. The BMTD model has a wide range of applications, including financial transactions, real time stock market data and accidents and events that occur irregularly with associated ”marks”.

Let us denote the data as $(x_t, y_t),\ t = 1, 2, \cdots, N$. (X_t, Y_t) is said to be generated from BMTD model if the conditional distribution of (X_t, Y_t) given the past can be written as

$$F\left(x_t, y_t|x^{t-1}, y^{t-1}\right) = \sum_{j=1}^{k} \pi_j F_j\left(x_t, y_t|x^{t-1}, y^{t-1}\right),$$
$$\sum_{j=1}^{k} \pi_j = 1, \quad \pi_j > 0, \quad j = 1, 2, \cdots, k.$$

Here $F_j(x_t, y_t|x^{t-1}, y^{t-1})$ is the joint distribution of (X_t, Y_t) conditional on $(X^{t-1}, Y^{t-1}) = (x^{t-1}, y^{t-1}) = ((x_{t-1}, y_{t-1}), (x_{t-2}, y_{t-2}), \cdots, (x_1, y_1))$ which represent the past information until time $t-1$; $F_j's$ are in a class of bivariate joint distributions.

Hassan and Lii introduced a general form of bivariate joint distribution for variable (X, Y). The joint density function for (X, Y) they defined has the following form

$$f_{X,Y}(x, y) = Cx^{\delta+\gamma+1/\phi-1}\left|\frac{y-\mu}{\beta}\right|^{\delta} e^{-x^{\alpha}(\lambda+|y-\mu|^{\phi}/\beta^{\phi})}, \quad x > 0, \quad -\infty < y < \infty \qquad (4.1)$$

where α, ϕ, δ and γ are all positive shape parameters; β and λ are positive scale parameters; and μ is a location parameter. The normalizing constant C is given by

$$C = \frac{1}{2\beta}\frac{\alpha\phi\lambda^{\delta/\alpha+\gamma/\alpha-\delta/\phi+1/(\alpha\phi)} - 1/\phi}{\Gamma\left(\frac{\delta+1}{\phi}\right)\Gamma\left(\frac{\delta}{\alpha}+\frac{\gamma}{\alpha}-\frac{\delta}{\phi}+\frac{1}{\alpha\phi}-\frac{1}{\phi}\right)}.$$

Restrict the ranges of x and y to $x > 0$ and $y > 0$ in (4.1) and put $\mu = 0$, the joint density function for (X, Y) they defined has the following form

$$f_{X,Y}(x, y) = Cx^{\delta+\gamma+1/\phi-1}y^{\delta}e^{-x^{\alpha}(\lambda+y^{\phi}/\beta^{\phi})} \quad x > 0, \ y > 0. \qquad (4.2)$$

where C is a constant which makes (4.2) a density function and C equals to

$$C = \frac{\alpha\phi\lambda^{\delta/\alpha+\gamma/\alpha-\delta/\phi+1/(\alpha\phi)-1/\phi}}{\beta^{\delta+1}\Gamma(\frac{\delta+1}{\phi})\Gamma(\frac{\delta}{\alpha}+\frac{\gamma}{\alpha}-\frac{\delta}{\phi}+\frac{1}{\alpha\phi}-\frac{1}{\phi})}.$$

For different value of α, ϕ, δ, it generates different kinds of bivariate joint distributions. For example,

1. Let $\alpha = \phi = 1$, $\delta = 0$, we will get Gamma-Pareto distribution like

$$f(x, y) = \frac{\lambda^{\gamma} x^{\gamma} e^{-x(\lambda + y/\beta)}}{\Gamma(\gamma)\beta}. \tag{4.3}$$

2. Let $\delta = 0, \phi = 2,$ and $\alpha = 1$, then the joint density is exponential-Cauchy type

$$f(x, y) = \frac{\lambda^{\gamma} x^{\gamma - 1/2} e^{-x(\lambda + (y-\mu)^2/\beta^2)}}{\Gamma(\gamma)\sqrt{\pi}\beta}. \tag{4.4}$$

Based on the joint distribution above, Hassan and Lii (2006) introduced the BMTD model which is the mixture of the joint distribution above and includes lag information. For example, for model (4.3), the mixture distribution is

$$f(x_t, y_t | x^{t-1}, y^{t-1}) = \sum_{j=1}^{k} \pi_j \frac{\lambda_j^{\gamma} x_t^{\gamma} e^{-x_t(\lambda_j + y_t/\beta_j)}}{\Gamma(\gamma)\beta_j} \tag{4.5}$$

and the corresponding re-parameterize of the parameters are

$$\lambda_j = \frac{1}{\theta_j x_{t-j} e^{-y_{t-j}}}.$$

For model (4.4), the mixture distribution is

$$f(x_t, y_t | x^{t-1}, y^{t-1}) = \sum_{j=1}^{k} \pi_j \frac{\lambda_j^{\gamma} x_t^{\gamma - 1/2} e^{-x_t(\lambda_j + (y_t - \mu_j)^2/\beta_j^2)}}{\Gamma(\gamma)\sqrt{\pi}\beta_j}. \tag{4.6}$$

Re-parametrization of the parameters can be taken as

$$\lambda_j = \frac{1}{\delta_j(1 + x_{t-j})},$$

$$\mu_j = \psi_j y_{t-j}.$$

4.2 Parameters Estimation with EM Algorithm

Hassan and Lii (2006) introduced EM algorithm for the BMTD model. It includes these steps

1. First set the number of component k as a fixed number;
2. Assign hidden variable Z as the indicator of the component that the observed data comes from;
3. E-step is to estimate the expectation value of Z;
4. M-step is to maximize the conditional likelihood function;
5. For different k, they get different models, then calculate BIC for these models;
6. The correct model is chosen as the one which maximized the BIC.

For example, look at the Gamma-Pareto distribution example, the joint distribution for this BMTD models is

$$f(x_t, y_t|x^{t-1}, y^{t-1}) = \sum_{j=1}^{k} \pi_j f_j = \sum_{j=1}^{k} \pi_j \frac{\lambda_j^{\gamma} x_t^{\gamma-1/2} e^{-x_t(\lambda_j + (y_t-\mu_j)^2/\beta_j^2)}}{\Gamma(\gamma)\sqrt{\pi}\beta_j}$$

where

$$\lambda_j = \frac{1}{\delta_j(1+x_{t-j})},$$
$$\mu_j = \psi_j y_{t-j}.$$

Suppose the observations of (X, Y) are represented by n observations given by

$(x_1, y_1), \cdots, (x_n, y_n)$. Let $Z = (Z_1, \cdots, Z_n)$ be the hidden variable where Z_t is a k-dimension indication vector and $Z_t = (Z_{t1}, \cdots, Z_{tk})$, where Z_{tj} is unity if observation (x_t, y_t) comes from component j and 0 otherwise. Also assume $Z's$ are independent of (X, Y) and $Z_t's$ are independent of each other. The conditional log-likelihood function can be written as

$$ll = \sum_{t=k+1}^{n} \sum_{j=1}^{k} z_{tj} \log \pi_j + \sum_{t=k+1}^{n} \sum_{j=1}^{k} z_{tj} \log f_j.$$

- E-Step: Suppose the parameters are known. The conditional expectation of $\tilde{z}_{tj}$ equals to

$$\tilde{z}_{tj} = \frac{\pi_j f_j(x_t, y_t | x^{t-j}, y^{t-j})}{\sum_{j=1}^{k} \pi_j f_j(x_t, y_t | x^{t-j}, y^{t-j})}.$$

- M-Step: Suppose $z_t's$ are known. Take partial derivatives with respect to the parameters, we have $\frac{\partial ll}{\beta_j} = 0$, $\frac{\partial ll}{\theta_j} = 0$, $\frac{\partial ll}{\pi_j} = 0$ for $j = 1, \cdots, k$. The estimates of parameters are

$$\hat{\alpha}_j = \frac{\sum_{t=k+1}^{n} \tilde{z}_{tj}}{\sum_{t=k+1}^{n} \sum_{j=1}^{k} \tilde{z}_{tj}},$$
$$\hat{\theta}_j = \frac{\sum_{t=k+1}^{n} \tilde{z}_{tj} x_t e^{-y_{t-j}}}{\gamma \sum_{t=k+1}^{n} \tilde{z}_{tj}},$$
$$\hat{\beta}_j = \frac{\sum_{t=k+1}^{n} \tilde{z}_{tj} x_t y_t}{\sum_{t=k+1}^{n} \tilde{z}_{tj}}.$$

After they got the estimation of the parameters, they plugged them into the likelihood function to calculate BIC. Finally the model is selected by which maximized the BIC.

4.3 Degeneracy of BMTD Models and Its Solution

To avoid confusing with parameter β_j in the inverse Gamma prior, we denote the β_j^2 in the mixture distribution as σ_j^2. From the general model of Bivariate distribution of (4.1) we see that if $\mu = y$ and $\sigma \to 0$ then the distribution will go to infinity since there is σ in the denominator in the constant C.

For more details, let's look at the exponential-Cauchy BMTD model (4.6).

The exponential part is $\exp\left\{-x_t\left(\frac{1}{\delta_j(1+x_{t-j})} + (y_t - \psi_j y_{t-j})^2/\sigma_j^2\right)\right\}$. We can see that if one of $\psi_j = y_t/y_{t-j}$ and $\sigma_j \to 0$, then the exponential part will be equal to a finite value $\exp\left\{-\frac{x_t}{\delta_j(1+x_{t-j})}\right\}$.Therefore, if $\sigma_j \to 0$ in the denominator, the product tends to be infinity. Under this condition we see that the joint density function will go to infinity.

As we have done in the previous section, we will use Bayesian method to solve this singularity problem.

If we assign a suitable prior for σ_j, such as Inverse Gamma, $IG(\alpha_j, \beta_j)$, then we can see the constant part containing σ_j is proportional to

$$
\begin{aligned}
&\frac{\lambda_j^\gamma x_t^{\gamma-1/2} e^{-x_t(\lambda_j + (y_t - \psi_j y_{t-j})^2/\sigma_j^2)}}{\Gamma(\gamma)\sqrt{\pi}\sigma_j} \times \left(\frac{1}{\sigma_j^2}\right)^{a_j} \exp\left\{-\frac{b_j}{\sigma_j^2}\right\} \\
&\propto \frac{\lambda_j^\gamma x_t^{\gamma-1/2}}{\Gamma(\gamma)\sqrt{\pi}} \frac{1}{\sigma_j} \exp\left\{-x_t\left(\lambda_j + \frac{(y_t - \psi_j y_{t-j})^2}{\sigma_j^2}\right)\right\} \times \left(\frac{1}{\sigma_j^2}\right)^{a_j} \exp\left\{-\frac{b_j}{\sigma_j^2}\right\} \\
&\propto \frac{\lambda_j^\gamma x_t^{\gamma-1/2}}{\Gamma(\gamma)\sqrt{\pi}} \exp\left\{-x_t\left(\lambda_j + \frac{(y_t - \psi_j y_{t-j})^2}{\sigma_j^2}\right)\right\} \times \left(\frac{1}{\sigma_j^2}\right)^{a_j+1/2} \exp\left\{-\frac{b_j}{\sigma_j^2}\right\}.
\end{aligned}
$$

Now from above we can see as $\psi_j = y_t/y_{t-j}$ and $\sigma_j^2 \to 0$, the part $\frac{\lambda_j^\gamma x_t^{\gamma-1/2}}{\Gamma(\gamma)\sqrt{\pi}}$ is a finite part. $\exp\left\{-x_t\left(\lambda_j + \frac{(y_t - \psi_j y_{t-j})^2}{\sigma_j^2}\right)\right\}$ will be zero since $\psi_j = y_t/y_{t-j}$. But the last part

$\left(\frac{1}{\sigma_j^2}\right)^{a_j+1/2} \exp\left\{-\frac{b_j}{\sigma_j^2}\right\}$ will be zero when $\sigma_j^2 \to 0$ because $\exp\left\{-\frac{b_j}{\sigma_j^2}\right\}$ will go to zero in exponential speed. So, totally the entire posterior will go to zero on the boundary of the parameter space. This helps to explain why Bayesian method can solve the problem of singularity.

In fact, we can find a general rule for the question of singularity. In the density function, if the exponential part containing random variable has a parameter in the denominator, and the normalizing constant part also contains that parameter in the denominator, then usually singularities will exist for this kind of density functions. The reason is if the parameter in the denominator (σ_j) goes to zero and the numerator in the exponential part goes to zero ($\mu_j \to x_i$) then the whole exponential part will go to a finite constant but the normalizing constant will be infinity since the parameter (σ_j) can converge to positive zero in the denominator.

4.4 Bayesian Method Study of BMTD Model

In this section, we will re-study the example that Lii and Hassan (2006) used in their paper under Bayesian framework. Here we will use the exponential-Cauchy Distribution model (4.6) as an example to study and compare the Bayesian and non-Bayesian methods in the simulation and parameter estimation.

Denote the data as $(x_t, y_t),\ t = 1, 2, \cdots, N$. (X_t, Y_t) is said to be generated from BMTD model with exponential-Cauchy distribution if the conditional density of (X_t, Y_t) given the past can be written as (to avoid confusing with parameter β_j in the inverse Gamma prior, we denote the β_j^2 in the mixture distribution (4.6) as σ_j^2)

$$f(x_t, y_t|x^{t-1}, y^{t-1}) = \sum_{j=1}^{k} \pi_j \frac{\lambda_j^\gamma x_t^{\gamma-1/2} e^{-x_t(\lambda_j + (y_t-\mu_j)^2/\sigma_j^2)}}{\Gamma(\gamma)\sqrt{\pi}\sigma_j}. \tag{4.7}$$

With the following re-parameterization to incorporate the lag information

$$\lambda_j = \frac{1}{\delta_i(1+x_{t-j})},$$
$$\mu_j = \theta_j y_{t-j}.$$

Based on this, we can write the distribution of (X_t, Y_t) given the previous information $(X^{t-1}, Y^{t-1}) = (X_1, \cdots, X_{t-1}, Y_1, \cdots, Y_{t-1})$ as follows

$$\begin{aligned} f(x_t, y_t|x^{t-1}, y^{t-1}) &= \sum_{j=1}^{k} \pi_j \frac{x_t^{1/2} e^{-x_t\left(\frac{1}{\delta_j(1+x_{t-j})} + \frac{(y_t-\theta_j y_{t-j})^2}{2\sigma_j^2}\right)}}{\delta_j(1+x_{t-j})\sqrt{2\pi}\sigma_j} \\ &= \sum_{j=1}^{k} \pi_j \frac{x_t^{1/2} e^{\frac{-x_t}{\delta_j(1+x_{t-j})}} e^{-x_t\frac{(y_t-\theta_j y_{t-j})^2}{2\sigma_j^2}}}{\delta_j(1+x_{t-j})\sqrt{2\pi}\sigma_j}. \end{aligned} \tag{4.8}$$

Suppose $\mathrm{Z}_1, \cdots, \mathrm{Z}_n$ are the hidden random variables where $\mathrm{Z_t} = (Z_{t1}, \cdots, Z_{tk})$ indicates the component that data comes from as discussed before. We also assume all $\mathrm{Z}'s$ are independent of the data (X, Y). Now we can write the (conditional) likelihood function of (x_t, Y_t) given the past as

$$\begin{aligned} L &\propto \prod_{t=k+1}^{n}\prod_{j=1}^{k} \left(\pi_j f(x_t, y_t|x^{t-1}, y^{t-1})\right)^{Z_{tj}} \\ &\propto \prod_{t=k+1}^{n}\prod_{j=1}^{k} \left(\pi_j \frac{x_t^{1/2} \exp\left\{\frac{-x_t}{\delta_j(1+x_{t-j})}\right\} \exp\left\{-x_t \frac{(y_t-\theta_j y_{t-j})^2}{2\sigma_j^2}\right\}}{\delta_j(1+x_{t-j})\sqrt{2\pi}\sigma_j}\right)^{Z_{tj}}. \end{aligned}$$

Since the singularity in the exponential-Cauchy model is caused by the σ_j, we choose non-informative prior distribution for the parameters δ_j and θ_j. We will choose Inverse Gamma $IG(\alpha_j, \beta_j)$ as the prior for σ_j^2. That is, the prior of $\frac{1}{\sigma_j^2}$ is

$$f\left(\frac{1}{\sigma_j^2}|\alpha_j, \beta_j\right) \propto \left(\frac{1}{\sigma_j^2}\right)^{\alpha_j} \exp\left\{-\beta_j \frac{1}{\sigma_j^2}\right\}. \tag{4.9}$$

Based on this, the (conditional) posterior distribution becomes

$$\begin{aligned} posterior \propto \prod_{t=k+1}^{n} \prod_{j=1}^{k} \left(\pi_j \frac{x_t^{1/2} \exp\left\{\frac{-x_t}{\delta_j(1+x_{t-j})}\right\} \exp\left\{-x_t \frac{(y_t-\theta_j y_{t-j})^2}{2\sigma_j^2}\right\}}{\delta_j(1+x_{t-j})\sqrt{2\pi}\sigma_j} \right)^{Z_{tj}} \\ \times \prod_{j=1}^{k} \frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)} \left(\frac{1}{\sigma_j^2}\right)^{\alpha_j} \exp\left\{-\frac{\beta_j}{\sigma_j^2}\right\}. \end{aligned}$$

Take log over the (conditional) posterior density, we get

$$\begin{aligned} ll &= \log(posterior) \\ &\propto \sum_{t=k+1}^{n} \sum_{j=1}^{k} z_{tj} \log \pi_j + \sum_{t=k+1}^{n} \sum_{j=1}^{k} z_{tj} \log f_j\left(x_t, y_t | x^{t-1}, y^{t-1}\right) \\ &\quad + \sum_{j=1}^{k} \log IG(\sigma_j^2|\alpha_j, \beta_j) \\ &\propto \sum_{t=k+1}^{n} \sum_{j=1}^{k} Z_{tj} \left(\log \pi_j - \frac{x_t}{\delta_j(1+x_{t-j})} - x_t \frac{(y_t-\theta_j y_{t-j})^2}{2\sigma_j^2} - \log \delta_j \right. \\ &\quad \left. -\frac{1}{2} \log \sigma_j^2 \right) + \sum_{j=1}^{k} \left(-\alpha_j \log \sigma_j^2 - \frac{\beta_j}{\sigma_j^2} \right). \end{aligned} \tag{4.10}$$

Next we use EM algorithm to get parameter estimation.

Use the same procedure as before we obtain

$$\tilde{z}_{tj} = \frac{\pi_j f_j\left(x_t, y_t | x^{t-1}, y^{t-1}\right)}{\sum_{j=1}^{k} \pi_j f_j\left(x_t, y_t | x^{t-1}, y^{t-1}\right)} \tag{4.11}$$

for $j = 1,\ 2, \cdots, k$.

$$\hat{\pi}_j = \frac{\sum_{t=k+1}^{n} \tilde{z}_{tj}}{\sum_{t=k+1}^{n} \sum_{j=1}^{k} \tilde{z}_{tj}}, \tag{4.12}$$

$$\hat{\delta}_j = \frac{\sum_{t=k+1}^{n} \frac{x_t}{1+x_{t-j}} \tilde{z}_{tj}}{\sum_{t=k+1}^{n} \tilde{z}_{tj}}. \tag{4.13}$$

$$\hat{\theta}_j = \frac{\sum_{t=k+1}^{n} \tilde{z}_{tj} x_t y_t y_{t-j}}{\sum_{t=k+1}^{n} \tilde{z}_{tj} x_t y_{t-j}^2}. \tag{4.14}$$

$$\begin{aligned} \hat{\sigma}_j^2 &= \frac{\sum_{t=k+1}^{n} \tilde{z}_{tj} x_t (y_t - \hat{\theta}_j y_{t-j})^2 + 2\beta_j}{\sum_{t=k+1}^{n} \tilde{z}_{tj} + 2\alpha_j} \\ &= \frac{\sum_{t=k+1}^{n} \tilde{z}_{tj} x_t (y_t - \hat{\theta}_j y_{t-j})^2}{\sum_{t=k+1}^{n} \tilde{z}_{tj} + 2\alpha_j} + \frac{2\beta_j}{\sum_{t=k+1}^{n} \tilde{z}_{tj} + 2\alpha_j}. \end{aligned} \tag{4.15}$$

The proof of (4.12), (4.13), (4.14) and (4.15) is in last section of this chapter.

From (4.13), (4.14) and (4.15) we see that the estimations of $\delta_j, \theta_j, j = 1, \cdots, k$ in Bayesian method is the same as they are by non-Bayesian method since we choose the non-informative prior for these parameters.

Now look at the estimation of σ_j^2 in Bayesian method. We see that as the sample size n increase, the summation of $\sum_{t=1}^{n} \tilde{z}_{tj}$ will also increase. If we get a large sample size n, the part $\sum_{t=1}^{n} \tilde{z}_{tj} + 2\alpha_j$ will be close to $\sum_{t=1}^{n} \tilde{z}_{tj}$ and $\frac{2\beta_j}{\sum_{t=1}^{n} \tilde{z}_{tj} + 2\alpha_j}$ will be close to zero since $\sum_{t=1}^{n} \tilde{z}_{tj}$ in the denominator is very large compared to the fixed β_j in the numerator. Also $\frac{\sum_{t=k+1}^{n} \tilde{z}_{tj} x_t (y_t - \hat{\theta}_j y_{t-j})^2}{\sum_{t=k+1}^{n} \tilde{z}_{tj} + 2\alpha_j}$ will converge to $\frac{\sum_{t=k+1}^{n} \tilde{z}_{tj} x_t (y_t - \hat{\theta}_j y_{t-j})^2}{\sum_{t=k+1}^{n} \tilde{z}_{tj}}$ which is the non-Bayesian estimator. That means, if we increase the sample size n, Bayesian estimator and non-Bayesian estimator will have less and less differences. But Bayesian method will assure that the estimator will not go to the boundary of parameter space which guarantees that a proper maximum of the posterior can be reached.

4.5 Simulation

Previously we have shown the superiorities of Bayesian method in the BMTD models. Now we will use simulations to compare them and verify the superiorities of Bayesian method. We use BMTD of exponential-Cauchy distribution with $k = 3$ mixture components. From the true model we sample $n = 200$ data points as our data. For the simulation, we first set the value of the number of components k to a fixed value. After we obtained the estimation of the parameters, we then select the model by comparing the different value of BIC corresponding to different values of k. For both methods, we repeat the simulation 100 times independently. The final estimation is the average of the repeats.

The true values of the parameters in the exponential-Cauchy BMTD model (4.8) are given below.

True Value

π	δ	θ	σ
0.1	0.4	0.3	0.1
0.7	0.7	0.3	1.0
0.2	0.9	-2.5	5.0

And the initial value of the EM algorithm are given below.

Initial Values

π	δ	θ	σ
0.15	0.30	0.2	0.3
0.60	0.78	0.4	0.5
0.25	0.80	-2.0	4.0

4.5.1 k=3 n=200 Non-Bayesian Method

First look at the non-Bayesian method. Of all 100 replications, there are 3 times the singularities occured. For the other simulations without the appearance of singularity, we take average of the 97 independent estimations. The results are given below.

Non-Bayesian Method Simulation Result

π	δ	θ	σ
0.1045049 (.1)	0.3886475 (.4)	0.2578528 (.3)	0.1273875 (.1)
0.6973006 (.7)	0.6517466 (.7)	0.2778409 (.3)	0.9172626 (1)
0.1981944 (.2)	0.8152921 (.9)	-2.3325705 (-2.5)	4.6463886 (5)

STD error of Non-Bayesian Method Simulation

π	δ	θ	σ
0.03378788	0.19998170	0.05456871	0.10639560
0.04503336	0.06966322	0.01192587	0.08676654
0.03234068	0.15417010	0.14134460	0.75175000

4.5.2 k=3 n=200 Bayesian Method

Next is the result of Bayesian simulation. As stated above, we choose non-informative priors for the parameters $\delta_j, \theta_j, j = 1, \cdots, k$. And the priors for σ_j^2 are Inverse Gamma (4.9) $IG(1,1)$ distribution. Of all 100 replications of simulation, there are no singularity for the Bayesian Method. The simulation results are given below.

Bayesian Method Simulation Result

π	δ	θ	σ
0.1097807 (.1)	0.3980461 (.4)	0.2703176 (.3)	0.117084 (.1)
0.7025879 (.7)	0.6790021 (.7)	0.2920027 (.3)	1.099523 (1)
0.1876315 (.2)	0.8992515 (.9)	-2.4013305 (-2.5)	4.914751 (5)

STD error of Bayesian Method Simulation

π	δ	θ	σ
0.03634200	0.14550780	0.04588224	0.06939043
0.03975061	0.07648589	0.01161006	0.06493199
0.03315683	0.15114260	0.17736150	0.59700200

Comparing the Bayesian and non-Bayesian simulations above, we see that for all 12 parameters estimation, Bayesian method has 9 of the 12 estimations closer to the true value than the non-Bayesian method. Also, look at the standard error of the 12 estimations, there are 8 out of 12 Bayesian method estimations perform better than the non-Bayesian method.

Another superiority of Bayesian method here is its flexibility. In the model, if we want to let the estimations throw more weight on the data rather than on the prior information, we can choose some suitable values in the priors. For example, if we choose small value for $\alpha_j,\ j = 1, \cdots, k$ and $\beta_j,\ j = 1, \cdots, k$, then Bayesian method estimator will be very close to non-Bayesian method. If we have strong information of the prior, then we can choose large value of α_j and $\beta_j,\ j = 1, \cdots, k$ to let the priors have more influence in the estimations.

4.6 Proof of (4.12) - (4.15)

From (4.10) we know the log posterior is

$$\begin{aligned}
ll &= \log(posterior) \\
&\propto \sum_{t=k+1}^{n}\sum_{j=1}^{k} z_{tj}\log\pi_j + \sum_{t=k+1}^{n}\sum_{j=1}^{k} z_{tj}\log f_j\left(x_t, y_t|x^{t-1}, y^{t-1}\right) \\
&\quad + \sum_{j=1}^{k}\log IG(\sigma_j^2|\alpha_j, \beta_j) \\
&\propto \sum_{t=k+1}^{n}\sum_{j=1}^{k} Z_{tj}\left(\log\pi_j - \frac{x_t}{\delta_j(1+x_{t-j})} - x_t\frac{(y_t-\theta_j y_{t-j})^2}{2\sigma_j^2} - \log\delta_j\right. \\
&\quad \left. -\frac{1}{2}\log\sigma_j^2\right) + \sum_{j=1}^{k}\left(-\alpha_j\log\sigma_j^2 - \frac{\beta_j}{\sigma_j^2}\right) \\
&\propto \sum_{t=k+1}^{n}\sum_{j=1}^{k} Z_{tj}\log\pi_j - \sum_{t=k+1}^{n}\sum_{j=1}^{k} Z_{tj}\left(\frac{x_t}{\delta_j(1+x_{t-j})} + \log\delta_j\right) \\
&\quad - \sum_{t=k+1}^{n}\sum_{j=1}^{k} Z_{tj}x_t\frac{(y_t-\theta_j y_{t-j})^2}{2\sigma_j^2} - \frac{1}{2}\sum_{t=k+1}^{n}\sum_{j=1}^{k} Z_{tj}\log\sigma_j^2 \\
&\quad + \sum_{j=1}^{k}\left(-\alpha_j\log\sigma_j^2 - \frac{\beta_j}{\sigma_j^2}\right).
\end{aligned}$$

Take partial difference of ll with respect to $\pi_j,\ \delta_j,\ \theta_j,\ \sigma_j^2$.

By setting $\dfrac{\partial ll}{\partial \pi_j} = 0$, for $j = 1,\ 2,\ \cdots,\ k-1$, we obtain

$$\sum_{t=k+1}^{n} z_{tj}/\pi_j - \sum_{t=k+1}^{n} z_{tk}/\pi_k = 0,$$

$$\sum_{t=k+1}^{n} z_{tj}\pi_k = \sum_{t=k+1}^{n} z_{tk}\pi_j,$$

$$\frac{\pi_j}{\pi_k} = \frac{\sum_{t=k+1}^{n} z_{tj}}{\sum_{t=k+1}^{n} z_{tk}} \quad ,\text{for } j = 1,\ 2,\ \cdots,\ k-1,$$

$$\hat{\pi}_j = \frac{\sum_{t=k+1}^{n} z_{tj}}{\sum_{t=k+1}^{n} \sum_{j=1^k} z_{tj}}.$$

By setting $\dfrac{\partial ll}{\partial \delta_j} = 0$, we obtain

$$\sum_{t=k+1}^{n} z_{tj} \left(\frac{x_t}{1+x_{t-j}} \frac{1}{\delta_j^2} - \frac{1}{\delta_j} \right) = 0,$$

$$\sum_{t=k+1}^{n} z_{tj} \frac{x_t}{1+x_{t-j}} \frac{1}{\delta_j^2} - \sum_{t=k+1}^{n} z_{tj} \frac{1}{\delta_j} = 0,$$

$$\sum_{t=k+1}^{n} z_{tj} \frac{x_t}{1+x_{t-j}} \frac{1}{\delta_j} - \sum_{t=k+1}^{n} z_{tj} = 0,$$

$$\hat{\delta}_j = \frac{\sum_{t=k+1}^{n} \frac{x_t}{1+x_{t-j}} z_{tj}}{\sum_{t=k+1}^{n} z_{tj}}.$$

By setting $\dfrac{\partial ll}{\partial \theta_j} = 0$, we obtain

$$\sum_{t=k+1}^{n} z_{tj} \frac{x_t}{2\sigma_j^2} 2(y_t - \theta_j y_{t-j}) y_{t-j} = 0,$$

$$\sum_{t=k+1}^{n} z_{tj} x_t (y_t - \theta_j y_{t-j}) y_{t-j} = 0,$$

$$\sum_{t=k+1}^{n} z_{tj} x_t y_t y_{t-j} - \theta_j \sum_{t=k+1}^{n} z_{tj} y_{t-j} y_{t-j} = 0,$$

$$\hat{\theta}_j = \frac{\sum_{t=k+1}^{n} z_{tj} x_t y_t y_{t-j}}{\sum_{t=k+1}^{n} z_{tj} y_{t-j}^2}.$$

By setting $\dfrac{\partial ll}{\partial \sigma_j^2} = 0,$ we obtain

$$\begin{aligned}
&\sum_{t=k+1}^{n} z_{tj}\left(\frac{x_t(y_t-\theta_j y_{t-j})^2}{2}\left(\frac{1}{\sigma_j^2}\right)^2 - \frac{1}{2}\frac{1}{\sigma_j^2}\right) - \alpha_j \frac{1}{\sigma_j^2} + \beta_j\left(\frac{1}{\sigma_j^2}\right)^2 = 0,\\
&\sum_{t=k+1}^{n} z_{tj}\left(x_t(y_t-\theta_j y_{t-j})^2 - \sigma_j^2\right) - 2\alpha_j\sigma_j^2 + 2\beta_j = 0,\\
&\sum_{t=k+1}^{n} z_{tj}x_t(y_t-\theta_j y_{t-j})^2 + 2\beta_j - \sigma_j^2\left(\sum_{t=k+1}^{n} z_{tj} + 2\alpha_j\right) = 0,\\
&\hat{\sigma}_j^2 = \frac{\sum_{t=k+1}^{n} z_{tj}x_t(y_t-\hat{\theta}_j y_{t-j})^2 + 2\beta_j}{\sum_{t=k+1}^{n} z_{tj} + 2\alpha_j}\\
&\quad = \frac{\sum_{t=k+1}^{n} z_{tj}x_t(y_t-\hat{\theta}_j y_{t-j})^2}{\sum_{t=k+1}^{n} z_{tj} + 2\alpha_j} + \frac{2\beta_j}{\sum_{t=k+1}^{n} z_{tj} + 2\alpha_j}.
\end{aligned}$$

We finished the proof.

Chapter 5

Consistency of Bayesian EM Parameter Estimation

5.1 Background of Consistency Study

In the previous two chapters we have shown for many MTD and BMTD models, the likelihood may not be bounded. Therefore maximum of likelihood estimation doesn't exist since the likelihood will go to infinity. Render and Walker (1984) showed the unboundedness of likelihood caused the failure of convergence of EM algorithm.

Wald (1949) and Chanda (1954) studied the consistency of maximum likelihood estimator. Most papers about consistency study use ideas from Wald's paper in 1949. Here we will briefly review Wald's work in consistency study.

Suppose $F(x, \theta)$ is the distribution for samples X, it's either discrete for all θ or is absolutely continuous for all θ. For any θ and for any positive value ρ let $f(x, \theta, \rho)$

be the supreme of $f(x,\theta^{'})$ with respect to $\theta^{'}$ when $|\theta-\theta^{'}| \leq \rho$. For any positive r, let $\varphi(x,r)$ be the supreme of $f(x,\theta)$ with respect to θ when $|\theta| > r$. Furthermore, let $f^{*}(x,\theta,\rho) = f(x,\theta,\rho)$ when $f(x,\theta,\rho) > 1$ and $= 1$ otherwise. Similarly $\varphi^{*}(x,r) = \varphi(x,r)$ when $\varphi(x,r) > 1$ and $= 1$ otherwise. By default, all the expectations in this chapter are with respect to X. The density should follow this assumption.

Assumption 1. *For sufficiently small ρ and for sufficiently large r the expected values of$\log f^{*}(x,\theta,\rho)$ (with respect to X) $\int_{-\infty}^{\infty}\log f^{*}(x,\theta,\rho)dF(x,\theta_0)$ and $\int_{-\infty}^{\infty}\log\varphi^{*}(x,r)dF(x,\theta_0)$ are finite where θ_0 denote the true parameter.*

Another assumption is that the integral of the absolute of log likelihood at the true parameter should exist, that is:

Assumption 2. *For the true parameter θ_0 we have*

$$\int_{-\infty}^{\infty} |\log f(x,\theta_0)| dF(x,\theta_0) < \infty \tag{5.1}$$

Based on these assumptions, Wald (1949) gave these two theorems:

Theorem 1. *For any compact subset S of the parameter space Ω and true parameter $\theta_0 \notin S$, then $P\left(\lim_{n\to\infty}\sup_{\theta\in S}\frac{L(\mathbf{X},\theta)}{L(\mathbf{X},\theta_0)} = 0\right) = 1$.*

From the theorem, if true parameter is not in that compact set, then the supreme of the ratio of likelihood in the set to the likelihood at the true parameter is almost surely zero.

Theorem 2. *If $\bar{\theta}(\mathbf{X})$ makes $\frac{L(\mathbf{X},\bar{\theta}(\mathbf{X}))}{L(\mathbf{X},\theta_0)} \geq c > 0$, then $P\left(\lim_{n\to\infty}\bar{\theta}(\mathbf{X}) \to \theta_0\right) = 1$.*

Since MLE maximize the likelihood, the likelihood ratio is at least 1. That is, if $c = 1$, the consistency of $\bar{\theta}(\mathbf{X})$ is obtained.

5.2 Consistency Study of MTD/BMTD Models

In this section we study the consistency of estimators for MTD and BMTD models with Bayesian method.

Gabriela Ciuperca, Andrea Ridolfi, Jerome Idier (2000) studied the consistency of penalized maximum likelihood estimator of normal mixtures. In their paper, they divide the area of $\sigma \in (0, \infty)$ into two parts: $\sigma \in (0, \eta)$ and $\sigma \in [\eta, \infty)$. Then prove separately for these two intervals. Here we extend the penalized estimators to Bayesian framework (penalized function can be treated as a special prior in Bayesian framework). Different from mixture models in their study, we consider the consistency of estimators for MTD/BMTD models. Then we use similar ideas to prove the consistency of estimators that maximize the posterior.

From previous sections we see that singularity is caused by σ_j which is a parameter of the denominator of the density. Here we mainly focus on this condition. We use Bayesian method and assign proper priors for the parameters which cause singularity. Here we will further prove such estimator is consistent.

Suppose the MTD/BMTD density for x_t has the following form

$$f(x_t|x^{t-1}, \theta) = \sum_{j=1}^{k} \pi_j f_j(x_t|x^{t-1}, \theta_j).$$

Here f_j is the density function and π_j is the weight of j^{th} component. $\theta_j = (\xi_j, \sigma_j)$ is the parameters of the j^{th} component. σ_j is the parameter in the j^{th} component that may cause degeneracy of the density. ξ_j is other general parameters non-related to singularity. $\Theta = \{\theta = (\xi, \sigma), k, \xi_1, \cdots, \xi_k; \sigma_1, \cdots, \sigma_k, k \geq 1, \sigma_j > 0\}$ is the parameter space.

Then the likelihood function is

$$L = f_n(x_n, \cdots, x_{k+1}, \theta | x^k) = \prod_{t=k+1}^{n} f(x_t | x^{t-1}, \theta) = \prod_{t=k+1}^{n} \sum_{j=1}^{k} \pi_j f_j(x_t | x^{t-1}, \theta_j).$$

Suppose the prior for σ_j^2 is $g(\sigma_j)$. It satisfies these following four conditions:

1. $\lim_{\sigma \to 0} \frac{1}{\sigma^n} g(\sigma) = 0$ for any n.

2. $g(\sigma)$ is a many-to-one mapping from $(0, \infty)$ to $(0, G]$, where $G = \sup g(\sigma)$.

3. g is increasing in an open interval $(0, s]$.

4. g is continuous differentiable on $(0, \infty)$.

So, the prior is $g(\sigma) = \prod_{j=1}^{k} g(\sigma_j)$. Then the posterior is

$$h_n = L \times g(\sigma) = \prod_{t=k+1}^{n} \sum_{j=1}^{k} \pi_j f_j(x_t | x^{t-1}, \theta_j) \prod_{j=1}^{k} g(\sigma_j).$$

For consistency, we extend the definition of h_n above as

$$h_n \left(x_n, \cdots, x_{k+1}, \theta | x^k \right) = \begin{cases} 0, & if \ \exists j, \ \ \sigma_j = 0, \\ f_n(x_n, \cdots, x_{k+1}, \theta | x^k) \prod_{j=1}^{k} g(\sigma_j), & if \ \ \sigma_j > 0, \forall j \end{cases} \tag{5.2}$$

Besides the assumptions on the prior function g, we need more assumptions for the density function as below:

Assumption 3. *We assume that the expectation of the absolute of the log of the density $f(x_t | x^{t-1}, \theta)$ with respect to X_t exists for all t, that is*

$$E \left| \log f(x_t | x^{t-1}, \theta) \right| < \infty. \tag{5.3}$$

Assumption 4. *We also assume the expectation of the absolute of log of the posterior exists with respect to X_{k+1} exists, that is*

$$E\left(\left|\log h_{k+1}\left(X_{k+1}, \theta|x^k\right)\right|\right) < \infty. \tag{5.4}$$

Here $h_{k+1}\left(X_{k+1}, \theta|x^k\right)$ is the function h in (5.2) with $n = k+1$.

To prove consistency, first we will prove some lemmas. Let us denote the true parameter as $\{\theta_0 = (k, \xi_0, \sigma_0) = (\xi_{01}, \cdots, \xi_{0k}; \sigma_{01}, \cdots, \sigma_{0k})\} \in \Theta$.

Lemma 1. *There exist $\eta > 0$ such that $\eta < \sigma_{0j},\ j = 1 \cdots k$ such that*

$$E\left(\log h_{k+1}\left(X, \theta|x^k\right)\right) < E\left(\log h_{k+1}\left(X, \theta_0|x^k\right)\right) \tag{5.5}$$

for any $\theta \in \bar{\Theta}$ with $\min\limits_{j=1\cdots k} \sigma_j \in [0, \eta)$. Here $\bar{\Theta}$ means Θ and its boundary $\partial\Theta$.

Proof. For any $\theta \in \bar{\Theta}$ we define

$$\nu = \log h_{k+1}(x_{k+1}, \theta|x^k) - \log h_{k+1}(x_{k+1}, \theta_0|x^k). \tag{5.6}$$

We will prove $E(\nu) < 0$. Given $\theta \in \Theta$ we can write

$$\begin{aligned} E(e^\nu) &= E_{x_{k+1}|x^k}\left(\frac{h_{k+1}(x_{k+1}, \theta|x^k)}{h_{k+1}(x_{k+1}, \theta_0|x^k)}\right) \\ &= \int_R h_{k+1}(x_{k+1}, \theta|x^k)\frac{\prod_{j=1}^k g(\sigma_j)}{\prod_{j=1}^k g(\sigma_{0j})} dx_{k+1} \\ &= \frac{\prod_{j=1}^k g(\sigma_j)}{\prod_{j=1}^k g(\sigma_{0j})}. \end{aligned} \tag{5.7}$$

We define function $\omega : (0, \infty) \to (0, \frac{1}{2}]$, i.e. $\omega(\delta) = \frac{g(\delta)}{2G}$, then

$$E(e^\nu) = \prod_{j=1}^k \frac{\omega(\sigma_j)}{\omega(\sigma_{0j})}. \tag{5.8}$$

We take δ such that $\omega(\delta) = \prod_{j=1}^{k} \omega(\sigma_{0j})$. The existence of $\delta \in (0, \infty)$ is granted by the many to one character of the function ω. In order to define η and to prove the inequality of (5.5), we have to consider two cases:

1. $\delta < s$ then we set $\eta = \delta$;

2. $\delta > s$ then if $\omega(\delta) < \omega(s)$, from one-to-one character of the function ω over $(0, s)$, there exists $\eta \in (0, s]$ s.t. $\omega(\eta) = \omega(\delta)$ else if $\omega(\delta) > \omega(s)$ we take $\eta = s$.

In both cases, because

$$\omega(\eta) \leq \omega(\delta) = \prod_{j=1}^{k} \omega(\sigma_{0j}) < \omega(\sigma_{0j}), \qquad \text{for } j = 1, \cdots, k,$$

we have

$$\omega(\eta) \leq \omega(\sigma_{0j}), \forall j = 1, \cdots, k. \tag{5.9}$$

When $\sigma_{0j} > s, j = 1, \cdots, k$ we straightly have $\eta < \sigma_{0j}, \forall j$. Otherwise when $\min_{j=1,\cdots,k} \sigma_{0j} < s$ from (5.9) we have $\eta < \sigma_{0j}, \forall j$ (Because $\omega(\cdot)$ is monotonicly increasing in $(0, s)$).

So, in both cases, we have $\eta < \sigma_{0j}, \forall j = 1, \cdots, k$.

If $\min_{j=1,\cdots,k} \sigma_j \in (0, \eta)$, by taking the definition of $\omega(\cdot)$ and assumption (3) on $g(\cdot)$ into account, we have

$$E(e^{\nu}) = \frac{\prod_j \omega(\sigma_j)}{\omega(\delta)} \leq \frac{\prod_j \omega(\sigma_j)}{\omega(\eta)} < \frac{\omega\left(\min_{j=1,\cdots,k} \sigma_j\right)}{\omega(\eta)} \leq 1.$$

If we consider the definition by extension of Θ (including the condition $\sigma_j = 0$, and for $\sigma_j = 0, \nu = -\infty$), we have $E(e^{\nu}) < 1, \forall \theta \in \Theta | \min_{j=1,\cdots,k} \sigma_j \in (0, \eta)$.

Since $E(\nu) < E(e^{\nu}) < 1$, the proof is done. □

For $\theta \in \Theta$, we define the following function

$$\begin{cases} \omega_{k+1}(x,\theta,\rho|x^k) = \sup\limits_{\theta':|\theta'-\theta|<\rho} h_{k+1}(x_{k+1},\theta'|x^k), \quad \rho > 0, \\ \\ \omega_n(x_n,\cdots,x_{k+1},\theta,\rho|x^k) = \sup\limits_{\theta':|\theta'-\theta|<\rho} h_n(x_n,\cdots,x_{k+1},\theta'|x^k), \quad \rho > 0. \end{cases}$$

Theorem 3. *Let S be a compact (closed) subset of $\bar{\Theta}$ such that $S = \{\theta \in \bar{\Theta} \mid \exists j \in \{1,\cdots,k\}$ suchthat $\sigma_j \in [0,\eta)\}$ and s.t. $\theta_0 \notin S$, then*

$$P\left(\lim_{n\to\infty}\sup_{\theta\in S}\frac{h_n\left(x_n,\cdots,x_{k+1};\theta|x^k\right)}{h_n\left(x_n,\cdots,x_{k+1};\theta_0|x^k\right)} = 0\right) = 1.$$

Proof. If we take the definition of h_n for $\sigma_k = 0$ into account, we may consider only the case $\min\sigma_j > 0$. By Lemma 1, for each point $\theta \in S$, we can associate a positive value ρ_θ such that

$$E\left(\log\omega_{k+1}\left(x_{k+1},\theta,\rho_\theta|x^k\right)\right) < E\left(\log h_{k+1}\left(x_{k+1},\theta_0|x^k\right)\right).$$

(This is because $E\left(\log\omega_{k+1}\left(x_{k+1},\theta,\rho_\theta|x^k\right)\right) = E\left(\log h_{k+1}\left(x_{k+1},\theta|x^k\right)\right)$ $< E\left(\log h_{k+1}\left(x_{k+1},\theta_0|x^k\right)\right)$).

Since S is compact, it can be covered by a finite number of open balls. Here, the theorem is proved if we can show that

$$P\left(\lim_{n\to\infty}\log h_n\left(x_n,\cdots,x_{k+1};\theta|x^k\right) - \log h_n\left(x_n,\cdots,x_{k+1};\theta_0|x^k\right) = -\infty\right) = 1.$$

Let $S(\theta,\rho_\theta)$ be the ball centered at θ with radius ρ_θ. Denote $\bar{S}(\theta,\rho_\theta)$ as $S(\theta,\rho_\theta)$ and its boundary.

Given n, there exists $\tilde{\theta}^{(n)} \in \bar{S}(\theta, \rho_\theta)$ such that

$$\log \frac{\omega_n\left(x_n, \cdots, x_{k+1}; \theta, \rho_\theta | x^k\right)}{h_n\left(x_n, \cdots, x_{k+1}; \theta_0 | x^k\right)} = \log \frac{h_n\left(x_n, \cdots, x_{k+1}; \tilde{\theta}^{(n)} | x^k\right)}{h_n\left(x_n, \cdots, x_{k+1}; \theta_0 | x^k\right)}. \tag{5.10}$$

(*if without prior $g(\sigma)$, $\tilde{\theta}^{(n)}$ doesn't exist since* $\max h_n$ *doesn't exist.*)

For $\tilde{\theta}^{(n)}$ such that $\exists j = 1, \cdots, k$ with $\tilde{\sigma_j}^{(n)} = 0$, then

$$\log \frac{\omega_n\left(x_n, \cdots, x_{k+1}; \theta, \rho_\theta | x^k\right)}{h_n\left(x_n, \cdots, x_{k+1}; \theta_0 | x^k\right)} = -\infty. \tag{5.11}$$

If $\tilde{\sigma_j}^{(n)} > 0, \forall j = 1, \cdots, k$, we have

$$\log \frac{h_n\left(x_n, \cdots, x_{k+1}; \tilde{\theta}^{(n)} | x^k\right)}{h_n\left(x_n, \cdots, x_{k+1}; \theta_0 | x^k\right)} = \sum_{t=k+2}^{n} \log \frac{f\left(x_t, \tilde{\theta}^{(n)} | x^{t-1}\right)}{f\left(x_t, \theta_0 | x^{t-1}\right)} + \log \frac{h_{k+1}\left(x_{k+1}, \tilde{\theta}^{(n)} | x^k\right)}{h_{k+1}\left(x_{k+1}, \theta_0 | x^k\right)}.$$

Let's analyze the two right terms of the previous equation separately.

Since h_n is continuous with respect to $\theta \in \bar{\Theta}$, if $\tilde{\theta}^{(n)}$ contains $\tilde{\sigma_j}^{(n)} \to 0$, then from (5.11) we have

$$\log \frac{\omega_n\left(x_n, \cdots, x_{k+1}; \tilde{\theta}^{(n)} | x^k\right)}{h_n\left(x_n, \cdots, x_{k+1}; \theta_0 | x^k\right)} = -\infty \quad \text{as} \quad n \to \infty. \tag{5.12}$$

In our notation $\tilde{\theta}^{(n)}$ is a vector and it contains $\tilde{\sigma_j}^{(n)}$. If this $\tilde{\sigma_j}^{(n)} \geq \sigma_j > 0, \forall j = 1, \cdots, k$, for any sample size $n \geq k+1$. we define

$$Z_t\left(\tilde{\theta}^{(n)}\right) = \frac{f\left(x_t; \tilde{\theta}^{(n)} | x^{t-1}\right)}{f\left(x_t; \theta_0 | x^{t-1}\right)}, \qquad t \geq k+2.$$

Since function f is continuous with respect to θ, we have

$$Z_t\left(\tilde{\theta}^{(n)}\right) \leq Z_t = \frac{f\left(x_t, \theta^{S(t)} | x^{t-1}\right)}{f\left(x_t, \theta_0 | x^{t-1}\right)},$$

with

$$\theta^{S(t)} = \arg \sup_{\theta' \in \bar{S}(\theta, \rho_\theta)} f\left(x_t, \theta' | x^{t-1}\right).$$

Since $Z_t = Z_t = \frac{f\left(x_t,\theta^{S(t)}|x^{t-1}\right)}{f(x_t,\theta_0|x^{t-1})}$, then the expectation of Z_t with respect to X_t is

$$E(Z_t) = \int_R \frac{f\left(x_t, \theta^{S(t)}|x^{t-1}\right)}{f\left(x_t, \theta_0|x^{t-1}\right)} f\left(x_t, \theta_0|x^{t-1}\right) dx_t = 1,$$

and we know $E(\log Z_t) < \log E(Z_t) = 0$, so

$$E(\log Z_t) < 0, \forall\ t = k+2, \cdots, n.$$

By strong law of large numbers, we have

$$\sum_{t=k+2}^{n} \log Z_t \to -\infty \quad a.s. \tag{5.13}$$

Let

$$Y = \log \frac{h_{k+1}\left(x_{k+1}; \tilde{\theta}^{(n)}|x^k\right)}{h_{k+1}\left(x_{k+1}, \theta_0|x^k\right)},$$

since $\tilde{\sigma_j}^{(n)} \in (0, \eta)$, so use Lemma 1 we have $E(Y) < 0$. And $E(Y) < 0$ means $P(Y = +\infty) = 0$. So from (5.13) and $P(Y = +\infty) = 0$, we have

$$\log \frac{\omega_n\left(x_n, \cdots, x_{k+1}; \theta, \rho_\theta|x^k\right)}{h_n\left(x_n, \cdots, x_{k+1}; \theta_0|x^k\right)} \xrightarrow[n\to\infty]{a.s.} -\infty, \tag{5.14}$$

From (5.10), (5.11), (5.12) and (5.14)

$$P\left(\lim_{n\to\infty} \sup_{\theta\in S} \frac{h_n\left(x_n, \cdots, x_{k+1}; \theta|x^k\right)}{h_n\left(x_n, \cdots, x_{k+1}; \theta_0|x^k\right)} = 0\right) = 1.$$

Theorem is proved. □

Theorem 4. *Let S be a compact (closed) subset of $\bar{\Theta}$ such that*

$S = \left\{\theta \in \bar{\Theta} \mid \exists j \in \{1, \cdots, k\}\ suchthat\ \sigma_j \in [\eta, +\infty)\right\}$ *and such that $\theta_0 \notin S$, then*

$$P\left(\lim_{n\to\infty} \sup_{\theta\in S} \frac{h_n\left(x_n, \cdots, x_{k+1}; \theta|x^k\right)}{h_n\left(x_n, \cdots, x_{k+1}; \theta_0|x^k\right)} = 0\right) = 1. \tag{5.15}$$

Proof. If $h_n(\cdot)$ in equation (5.15) is likelihood function, Render (1981, Theorem 3) proved if the expectation of the absolute value of log likelihood with respect to X existed, then (5.15) was correct. We just need to verify that the posterior here meets the requirements (the expectation of the absolute value of log posterior with respect to X existed) of Render (1981).

Under the assumption that the expectation of the absolute of the density $f(x_t|x^{t-1}, \theta)$ with respect to X exists, that is:

$$E\left|\log f(x_t|x^{t-1}, \theta)\right| < \infty.$$

Also, under the assumption that $g(\sigma) < G$, we know that $\prod_{j=1}^{k} g(\sigma_j) < \infty$. So

$$\begin{aligned} & E\left|\log(f(x_t|x^{t-1}, \theta)\prod_{j=1}^{k} g(\sigma_j))\right| \\ & = E\left|\log(f(x_t|x^{t-1}, \theta))\right|\prod_{j=1}^{k} g(\sigma_j) < \infty. \end{aligned} \tag{5.16}$$

So our posterior here satisfies the assumption of Render (1981). Therefore the theorem is proved. □

Theorem 5. *Let $\bar{\theta}_n = \bar{\theta}_n(x_1, \cdots, x_n) \in \bar{\Theta}$ be a function of $x_1, \cdots, x_n$ such that*

$$\frac{h_n\left(x_n, \cdots, x_{k+1}; \bar{\theta}_n|x^k\right)}{h_n\left(x_n, \cdots, x_{k+1}; \theta_0|x^k\right)} \geq \rho > 0, \quad \forall x_1, \cdots, x_n, \quad \forall n, \tag{5.17}$$

then

$$P\left(\lim_{n\to\infty} \bar{\theta}_n = \theta_0\right) = 1.$$

Proof. It's sufficient to prove that for any fixed $\epsilon > 0$, we have

$$P\left(\lim_{n\to\infty} \bar{\theta}_n = \bar{\theta} \mid \parallel \bar{\theta} - \theta_0 \parallel \leq \epsilon\right) = 1.$$

From Render (1981, Theorem 5), the estimator is strongly consistent over $[\eta, \infty)$, the only condition that needs consideration is $\min \sigma_j \in [0, \eta)$.

If $\parallel \bar{\theta} - \theta_0 \parallel \geq \epsilon$, then $\bar{\theta} \in \{\theta : \parallel \theta - \theta_0 \parallel \geq \epsilon\}$. Denote the supremum of $h_n\left(x_n, \cdots, x_{k+1}; \theta | x^k\right)$ on $\{\theta : \parallel \theta - \theta_0 \parallel \geq \epsilon\}$ with respect to θ by $\sup\limits_{\|\theta-\theta_0\|\geq\epsilon} h_n\left(x_n, \cdots, x_{k+1}; \theta | x^k\right)$. So we have

$$\sup_{\|\theta-\theta_0\|\geq\epsilon} h_n\left(x_n, \cdots, x_{k+1}; \theta | x^k\right) \geq h_n\left(x_n, \cdots, x_{k+1}; \bar{\theta} | x^k\right).$$

Because $\lim_{n\to\infty} \bar{\theta}_n = \bar{\theta}$, from continuity of function $h_n\left(x_n, \cdots, x_{k+1}; \theta | x^k\right)$, we have

$$h_n\left(x_n, \cdots, x_{k+1}; \bar{\theta} | x^k\right) = h_n\left(x_n, \cdots, x_{k+1}; \bar{\theta}_n | x^k\right)$$

for sufficiently large n.

So, if $\parallel \bar{\theta} - \theta_0 \parallel \geq \epsilon$, we would have

$$\sup_{\|\theta-\theta_0\|\geq\epsilon} h_n\left(x_n, \cdots, x_{k+1}; \theta | x^k\right) \geq h_n\left(x_n, \cdots, x_{k+1}; \bar{\theta}_n | x^k\right). \tag{5.18}$$

Because of (5.17) and (5.18) we have

$$\sup_{\|\theta-\theta_0\|\geq\epsilon} \frac{h_n\left(x_n, \cdots, x_{k+1}; \theta | x^k\right)}{h_n\left(x_n, \cdots, x_{k+1}; \theta_0 | x^k\right)} \geq \rho > 0, \quad \text{for all } n \geq k+1. \tag{5.19}$$

However, according to Theorem 3, this is an event with probability zero. That is, it's probability one that all limit points $\bar{\theta}$ of $\bar{\theta}_n$ satisfy the inequality $\parallel \bar{\theta} - \theta_0 \parallel \leq \epsilon$. □

Since $\bar{\theta}_n$ maximizes the posterior density, if we set $\rho = 1$, the condition (5.17) is satisfied. So we get the consistency of $\bar{\theta}_n$.

Chapter 6

Birth-Death Process Method for MTD Normal

In Chapter 3 and Chapter 4 we have shown how to use EM algorithm to maximize the posterior. Under this framework, we first fix k then use BIC for model selection. Now in this chapter we will show another method in which k is also treated as a (discrete) random variable. We assign a prior for k and other parameters in the density. Rather than using EM algorithm to maximize the posterior, here we use Markov Chain Monte Carlo (MCMC) to sample from the posterior. The estimation of parameters is based on these samples from MCMC. The method of MCMC sampling here is called Birth-Death process which was introduced by Matthew Stephens (2000). He used this method to study normal mixture models. Here we will extend this method to Normal Mixture Transitions Distributions in which the re-parameterization contains build-in lag information.

6.1 Background Introduction

Richardson and Green (1997) present a method of performing a Bayesian analysis of data from a finite mixture distribution with an unknown number of components. Their method is an MCMC approach which makes use of the reversible jump method described by Green (1995). Matthew Stephens (2000) gave another approach to study the mixture model using Bayesian method. After getting the posterior distribution, a Markov Birth-Death process with Gibbs sampler was created to sample from the posterior distribution. This method is easier to implement than the reversible jump method (Stephens 2000) and it can be used if the data is more than one dimension. All the parameters including number of components are assigned priors and we sample from the posterior distribution. So this method is also called fully Bayesian method.

Stephens (2000) introduced a method of constructing an ergodic Markov chain with appropriate stationary distribution, when the number of components k is considered unknown. The method is based on the construction of a continuous time Markov Birth-Death process as described by Preston (1976) with the appropriate stationary distribution. In order to apply these MCMC methods to the mixture model context, they view the parameters of the model as a marked point process, with the point representing a component of the mixture. The MCMC scheme allows the number of components to vary by allowing new components to be born and existing components to die. These Births and Deaths occur in continuous time, and the relative rates at which they occur determine the stationary distribution of the process. They use this Birth-Death scheme to construct an easily simulated process, in which Births occur at a constant rate from the prior, and Deaths occur at a rate which is very low for components which are critical in explaining the data, and very high for components which do not help to explain the data. The accept-reject mechanism allows both good and bad Births to occur, but reverses bad Births very quickly through very quick Deaths.

6.2 MCMC Method and Gibbs Sampling

A Markov Chain in discrete time and general state space E is a sequence of random variables $(\Theta^0, \Theta^1, \cdots)$ with $\Theta^t \in E$, which obeys Markov Property in time. That is, given the current state $\Theta^t (t \geq 0)$ the distribution of the next state Θ^{t+1} is independent of the previous history of the chain, $(\Theta^0, \Theta^1, \cdots, \Theta^{t-1})$.

A Markov chain is said to be stationary (invariant) with stationary distribution π if $\Theta^t \sim \pi$ then $\Theta^{t+1} \sim \pi$ also.

MCMC methods rely on the construction of stationary Markov chain and construction of such a chain is often straightforward. An algorithm which has found wide application and is particularly suited to the mixture model is the Gibbs sampler.

Gibbs Sampler: The Gibbs sampler (Geman 1984) is a special MCMC scheme. It gives a method of constructing a Markov chain with a given stationary distribution $\pi(\theta|x^n)$ which is the posterior of θ given data $x^n = (x_1, \cdots, x_n)$. Suppose the random variable Θ can be decomposed into d components, $\Theta = (\Theta_1, \cdots, \Theta_d)$. Suppose we cannot sample directly from $\pi(\theta|x^n) = \pi((\theta_1, \cdots, \theta_d)|x^n)$ but can sample directly from the full conditional distributions

$$\pi(\theta_1|\theta_2, \cdots, \theta_d, x^n), \cdots, \pi(\theta_d|\theta_1, \cdots, \theta_{d-1}, x^n).$$

Here we describe a type of Gibbs Sampler strategy that is widely used in practice.

Systematic-Scan Gibbs Sampler. Let $\theta^{(t)} = (\theta_1^{(t)}, \cdots, \theta_d^{(t)})$ for iteration t, then to simulate a value for $\theta^{(t+1)}$ in the following d steps:

- Step 1: Sample θ_1^{t+1} from $\pi(\theta_1|\theta_2^{(t)}, \cdots, \theta_d^{(t)}, x^n)$;
- Step 2: Sample θ_2^{t+1} from $\pi(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, \cdots, \theta_d^{(t)}, x^n)$;

  $\cdots$
- Step d: Sample θ_d^{t+1} from $\pi(\theta_d|\theta_1^{(t+1)}, \cdots, \theta_{d-1}^{(t+1)}, x^n)$;

Later we will use this Gibbs sampler to sample the parameters from the posterior.

6.3 Hierarchical Model Expression of MTD Model

Here we consider the MCMC method for the normal mixture transition distribution (MTD) model which was introduced by Le, Martin and Raftery (1996). First, consider the normal mixture transition distribution model:

$$f\left(x_t|x^{t-1}\right) = \sum_{j=1}^{k} \pi_j \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{\left(x_t - \theta_j x_{t-j}\right)^2}{2\sigma_j^2}}.$$

Suppose the prior distribution for parameters $(k, \boldsymbol{\pi}, \boldsymbol{\phi})$ given parameters η is $r\left(k, (\pi_1, \cdots, \pi_k), (\phi_1 = (\theta_1, \sigma_1^2), \cdots, \phi_k = (\theta_k, \sigma_k^2))|\eta\right)$ and it is exchangeable. That is

$$r\left(k, (\pi_1, \cdots, \pi_k), (\phi_1, \cdots, \phi_k)\right) = r\left(k, (\pi_{p_1}, \cdots, \pi_{p_k}), (\phi_{p_1}, \cdots, \phi_{p_k})\right)$$

for all permutations of p of $(p_1, \cdots, p_k)$. Let $\mathscr{U}^{k-1}$ denote the uniform distribution on the simplex

$$\mathscr{S}_{k-1} = \{(\pi_1, \cdots, \pi_{k-1}) : \pi_1, \cdots, \pi_{k-1} \geq 0 \text{ and } \pi_1 + \cdots + \pi_{k-1} \leq 1\}.$$

Let Φ denote the parameter space for ϕ_i (that is, $\phi_i \in \Phi$); let ν be the measure on Φ. As a simple example of the prior, suppose given k, η, $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$ are conditionally independent, and $\phi_1, \cdots, \phi_k$ are also independently identically distributed with density function $p(\phi_i|\eta)$ for ϕ_i. $\boldsymbol{\pi}$ has uniform distribution on the simplex $\mathscr{U}^{k-1}$. Now the prior becomes

$$r(k, \boldsymbol{\pi}, \boldsymbol{\phi}) = p(k|\eta)p(\phi_1|\eta)\cdots p(\phi_k|\eta). \tag{6.1}$$

Here η is a known constant vector for the priors. The conditional likelihood function is

$$\begin{aligned} L(k, \boldsymbol{\pi}, \boldsymbol{\phi}) &= p(x^n|k, \boldsymbol{\pi}, \boldsymbol{\phi}) \\ &= \prod_{i=k+1}^{n} \left(\sum_{j=1}^{k} \pi_j f(x_i|\phi_j, x^{i-1}, y^{i-1}) \right). \end{aligned} \tag{6.2}$$

From (6.1) and (6.2), it's easy to get the posterior distribution

$$p(k, \boldsymbol{\pi}, \boldsymbol{\phi}|x^n \eta) \propto L(k, \boldsymbol{\pi}, \boldsymbol{\phi}) r(k, \boldsymbol{\pi}, \boldsymbol{\phi}). \tag{6.3}$$

What we need to do next is to sample from the posterior above. We can consider any set of k parameter values $\{(\pi_1, \phi_1), \cdots, (\pi_k, \phi_k)\}$ as a set of k points in $[0,1] \times \Phi$, with the constraint that $\pi_1 + \cdots + \pi_k = 1$. Thus the posterior distribution $p(k, \boldsymbol{\pi}, \boldsymbol{\phi}|x^n, \eta)$ can be treated as a suitably constrained distribution of points in $[0,1] \times \Phi$. Stephens (2000) stated that these points from the posterior distribution could be treated as a point process on $[0,1] \times \Phi$, in other words, a marked point process on Φ with each ϕ_i having an associated marks π_i with the constraint that the sum of marks is unity. Based on this, Ripley (1977) and Stephens (2000) stated that a Markov Birth-Death process could be constructed from the stationary posterior distribution. Next we will describe in detailed about the Birth-Death process.

6.4 Birth-Death Process

Ripley (1977), Stephens (2000) construct a continuous time Markov Birth-Death process with stationary distribution $p(k, \boldsymbol{\pi}, \boldsymbol{\phi}|x^n, \eta)$. In the similar way, we can construct a continuous time Markov Birth-Death process with stationary distribution $p(k, \boldsymbol{\pi}, \boldsymbol{\phi}|x^n, \eta)$ with η kept fixed.

Birth-Death for the components of a mixture model Let Ω_k denote the parameter space of the mixture model with k components and $\Omega = \cup_{k>1}\Omega_k$. Writing $\omega = \{(\pi_1, \phi_1), \cdots, (\pi_k, \phi_k)\} \in \Omega_k$ to represent the parameters of the model.

Birth: If at time t the process is at $\omega = \{(\pi_1, \phi_1), \cdots, (\pi_k, \phi_k)\} \in \Omega_k$ and if a Birth occur at $(\pi, \phi) \in [0,1] \times \Phi$, then the process jumps to

$$\omega \cup (\pi, \phi) := \{(\pi_1(1-\pi), \phi_1), \cdots, (\pi_k(1-\pi), \phi_k), (\pi, \phi)\} \in \Omega_{k+1}.$$

Death: If at time t the process is at $\omega = \{(\pi_1, \phi_1), \cdots, (\pi_k, \phi_k)\} \in \Omega_k$ and if a Death occur at $(\pi_k, \phi_k) \in \omega$, then the process jumps to

$$\omega \backslash (\pi, \phi) := \left\{ \left(\frac{\pi_1}{1-\pi_k}, \phi_1 \right), \cdots, \left(\frac{\pi_{k-1}}{1-\pi_k}, \phi_{k-1} \right) \right\}.$$

Thus a Birth increases the number of components by one, while a Death decreases the number of components by one. These definitions have been chosen so that Birth and Death are inverse operations to each other and the constraint $\pi_1 + \cdots + \pi_k = 1$ remains satisfied after Birth and Death.

More details, when the process is at $\omega = \{(\pi_1, \boldsymbol{\phi}_1), \cdots, (\pi_k, \boldsymbol{\phi}_k)\} \in \Omega_k$, let Birth and Death occur as independent Poisson process as follows:

- *Birth:*
  - Currently the process is at $\omega = \{(\pi_1, \boldsymbol{\phi}_1), \cdots, (\pi_k, \boldsymbol{\phi}_k)\}$;
  - Birth occurs at rate $\beta(\omega)$;
  - If a Birth occurs, then the new parameters are

    $\omega' = \{(\pi_1(1-\pi), \boldsymbol{\phi}_1), \cdots, (\pi_k(1-\pi), \boldsymbol{\phi}_k), (\pi, \boldsymbol{\phi})\} \in \Omega_{k+1}$;
  - New parameter $(\pi, \boldsymbol{\phi})$ is sampled from density function $b(\omega; (\pi, \boldsymbol{\phi}))$.

- *Death:*
  - Currently the process is at $\omega = \{(\pi_1, \boldsymbol{\phi}_1), \cdots, (\pi_k, \boldsymbol{\phi}_k)\}$;
  - Death happens with rate $\delta(\omega) = d\left(\omega \backslash (\pi_k, \boldsymbol{\phi}_k); (\pi_k, \boldsymbol{\phi}_k)\right)$; Here $\omega \backslash (\pi_k, \boldsymbol{\phi}_k)$ means ω excludes $(\pi_k, \boldsymbol{\phi}_k)$;
  - If a Death happens, the last point $(\pi_k, \boldsymbol{\phi}_k)$ dies and the new parameters are

$$\left\{ \left(\frac{\pi_1}{1-\pi_k}, \boldsymbol{\phi}_1 \right), \cdots, \left(\frac{\pi_{k-1}}{1-\pi_k}, \boldsymbol{\phi}_{k-1} \right) \right\} \in \Omega_{k-1}. \tag{6.4}$$

Since Birth and Death occur as independent Poisson process. We know the combined rate for Poisson process is the sum of rate for Birth and Death. The corresponding waiting time for a jump follows exponential distribution with mean equal to the inverse of the combined rate. So the time to the next Birth/Death event is then exponentially distributed, with mean $1/(\beta(\omega)+\delta(\omega))$, and it will be a Birth with probability $\beta(\omega)/(\beta(\omega)+\delta(\omega))$, and a Death with probability $\delta_(\omega)/(\beta(\omega)+\delta(\omega))$.

The following theorem gives sufficient conditions on b and d to make the Birth-Death process having stationary distribution $p(k,\boldsymbol{\pi},\boldsymbol{\phi}|x^n,\boldsymbol{\eta})$. (Stephens, 2000).

Theorem 6. *Assuming the general prior on $(k,\boldsymbol{\pi},\boldsymbol{\phi})$ given in (6.1) and the corresponding posterior is given in (6.3), the Birth-Death process defined above has stationary distribution $p(k,\boldsymbol{\pi},\boldsymbol{\phi}|x^n,\boldsymbol{\eta})$, provided b and d satisfy*

$$\begin{aligned}(k+1)d(\omega;(\pi,\boldsymbol{\phi}))r(\omega\cup(\pi,\boldsymbol{\phi})|\boldsymbol{\eta})L(\omega\cup(\pi,\boldsymbol{\phi}))k(1-\pi)^{k-1}&\\ =\beta(\omega)b(\omega;(\pi,\boldsymbol{\phi}))r(\omega|\boldsymbol{\eta})L(\omega)&\end{aligned}\tag{6.5}$$

for all $\omega\in\Omega_k$ and $(\pi,\phi)\in[0,1]\times\Phi$.

We can explain (6.5) in this way: for any k, the left part of equation (6.5) is like the death rate from $k+1$ mixtures to k mixtures times the likelihood of $k+1$ components mixture. We roughly treat it as the total rate from $k+1$ mixtures to k mixtures. The right part is the birth rate from k mixtures to $k+1$ mixtures times the likelihood of k mixture components, which can be roughly treated as the total rate from k mixtures to $k+1$ mixtures. The equation means that the rate from $k+1$ to k equal to the rate from k to $k+1$.

From this theorem, we choose $b(\omega; (\pi, \boldsymbol{\phi}))$ and $d\,(\omega\backslash(\pi_k, \boldsymbol{\phi}_k); (\pi_k, \boldsymbol{\phi}_k))$ as

$$b(\omega; (\pi, \boldsymbol{\phi})) = k(1-\pi)^{k-1} p(\boldsymbol{\phi}|\boldsymbol{\eta}), \tag{6.6}$$

$$d\,(\omega\backslash(\pi_k, \boldsymbol{\phi}_k); (\pi_k, \boldsymbol{\phi}_k)) = \beta(\omega) \frac{L(\omega\backslash(\pi_k, \boldsymbol{\phi}_k))}{L(\omega)} \frac{p(k-1|\lambda)}{kp(k|\lambda)} \tag{6.7}$$

to make the process from the posterior distribution stationary.

6.5 General Introduction of Birth-Death Process Algorithm

For the mixture model, k is the number of mixture components (mixtures) in the model.

At time t, number of mixtures is k; the corresponding parameters are

$$(\pi, \phi)_{(t)} = ((\pi_1, \cdots, \pi_k), (\theta_1, \cdots, \theta_k), (\sigma_1^2, \cdots, \sigma_k^2))_{(t)}.$$

When a jump (a birth or a death) happens at time $t+1$, a new component may be born or an existing component may die. This restricts our sample because the number of mixtures is restraint to change by 1 every time. This is not efficient since every time we can only increase or decrease the mixture component by one. To avoid this problem, we will let the Birth-Death run a given time t_0, if the cumulative running time is less than t_0, we will not record the jump result. If the cumulative running time is greater than t_0, we will record the result. Suppose this Birth-Death process runs t_0 time, then we get a new combination of the mixtures with k' components, denote as

$$(\pi, \phi)_{(t+1)} = ((\pi_1, \cdots, \pi_{k'}), (\theta_1, \cdots, \theta_{k'}), (\sigma_1^2, \cdots, \sigma_{k'}^2))_{(t+1)}. \tag{6.8}$$

Based on this, we can give the following algorithm.

Algorithm 1. *Beginning from* $\omega^{(0)} = (k, \pi, \phi)_{(0)}$ *from the prior. The state from* $\omega^{(t)} = (k, \pi, \phi)_{(t)}$ *to* $\omega^{(t+1)} = (k, \pi, \phi)_{(t+1)}$ *is as follows:*

1. *First we fix the* $\boldsymbol{\eta}$ *for the priors. Let the Birth rate* $\beta(\omega) = \lambda$*; calculate Death rate* $\delta(\omega) = d\left(\omega \backslash (\pi_k, \boldsymbol{\phi}_k); (\pi_k, \boldsymbol{\phi}_k)\right)$ *from (6.7);*

2. *Simulate the waiting time to next jump from exponential distribution with mean* $1/(\beta(\omega) + \delta(\omega))$*;*

3. *Simulate the type of jump: Birth or Death with respective probabilities*

$$Pr(Birth) = \frac{\beta(\omega)}{\beta(\omega) + \delta(\omega)}, \qquad Pr(Death) = \frac{\delta(\omega)}{\beta(\omega) + \delta(\omega)}.$$

4. *Update* ω *to reflect the Birth or Death as introduced before;*

5. *If waiting time* $< t_0$*, return to step 1 until the cumulative waiting time* $> t_0$*. Record updated* $(k, \boldsymbol{\pi}, \boldsymbol{\phi})$ *as* $(k^{(t)'}, \pi^{(t)'}, \phi^{(t)'})$*. Set* $k^{(t+1)} = k^{(t)'}$*.*

6. *Sample* $Z_{ij}^{(t+1)}$ *from the posterior distribution of* $P\left(Z_{ij}^{(t+1)} = 1 | k^{(t+1)}, \pi^{(t)'}, \boldsymbol{\phi}^{(t)'}, \boldsymbol{\eta}^{(t)}, x^n\right)$*, here* $j = 1, \cdots, k^{(t+1)}$*;*

7. *Sample* $\boldsymbol{\eta}^{(t+1)}$ *from the posterior* $f\left(\boldsymbol{\eta} | k^{(t+1)}, \pi^{(t)'}, \boldsymbol{\phi}^{(t)'}, (\boldsymbol{z})^{(t+1)}, x^n\right)$*;*

8. *Sample* $\pi^{(t+1)}$ *from* $p\left(\pi_1, \cdots, \pi_{k^{(t+1)}} | k^{(t+1)}, \boldsymbol{\phi}^{(t)'}, \boldsymbol{\eta}^{(t+1)}, (\boldsymbol{z})^{(t+1)}, x^n\right)$*;*

9. *Sample* $\theta_j^{(t+1)}, (\sigma_j^2)^{(t+1)}$ *from their posterior distribution conditional on the updated* $k^{(t+1)}, \pi^{(t+1)}, \boldsymbol{\eta}^{(t+1)}, \boldsymbol{z}^{(t+1)}, x^n$*.*

note: here $\boldsymbol{z}^{(t+1)}$ *gives the partition of data* x^n*; the* j^{th} *partition is* $C_j^{(t+1)}$ *and the size of* $C_j^{(t+1)}$ *is* $n_j^{(t+1)}$ *with* $n_j^{(t+1)} = \#\left\{i : x_i \in C_j^{(t+1)}\right\}, i = 1, \cdots, n,\ j = 1, \cdots, k^{(t+1)}$.

Now we get the updated

$\omega^{(t+1)} = (k, \pi, \phi)_{(t+1)} = \left(k^{(t+1)}, \pi^{(t+1)}, \boldsymbol{\phi}^{(t+1)}\right).$

If necessary, we can also assign priors for other parameters in the model, for example, we can assign priors for μ_0, σ_0^2 in the priors for θ.

The algorithm above requires to specify the value of a Birth-rate λ_b and the value of time t_0. From equation (6.5) we can see if we double λ_b, then the death rate will also be doubled; so we are free to set $t_0 = 1$ and only need to fix λ_b. In our examples we set $\lambda_b = \lambda$ (the parameter of the Poisson prior), which gives a convenient form of the Death rates as a likelihood ratio which does not depend on λ.

6.5.1 Parameter Estimation

We run the Birth-Death process for $M+N$ times, we can get the parameter set: $\omega^1, \cdots, \omega^M, \omega^{M+1}, \cdots, \omega^{M+N}$, (here $\omega^i = ((\pi_1, \phi_1), \cdots, (\pi_{k_i}, \phi_{k_i}))$ has k_i components); Throwing away the initial M points as the burning-in, the rest $\omega^{(M+1)}, \cdots, \omega^{(M+N)}$ are used as the sample. Then we can estimate the probability of number of components by an appropriate sample path average. That is

$$\begin{aligned} Pr(k=i|x^n) &= E(I(k=i)|x^n) \\ &\approx \frac{1}{N}\sum_{j=M+1}^{M+N} I(k_j=i) \\ &= \frac{1}{N}\#\{j:k_j=i\}. \end{aligned} \quad (6.9)$$

Suppose the estimated number of components $k = k_0$. A set that collects all indices of partition with $k_j = k_0$ is defined as $\mathcal{N} = \{j : k_j = k_0\}$. Parameter estimation could be defined as the average of the parameters over the set $\mathcal{N}$. That is, $\omega = \frac{1}{N_0}\sum_{j\in\mathcal{N}}\omega^j$ where N_0 is the size of $\mathcal{N}$.

6.6 Posterior Density of MTD Normal

Consider our MTD normal model

$$f\left(x_t|x^{t-1}\right) = \sum_{j=1}^{k}\pi_j\frac{1}{\sqrt{2\pi}\sigma_j}e^{-\frac{\left(x_t-\theta_j x_{t-j}\right)^2}{2\sigma_j^2}}.$$

As before, we introduce the hidden variable z_{tj}; $z_{tj} = 1$ if x_t is from component j and $z_{tj} = 0$ if x_t is from other component. That is

$$\begin{aligned} &F(x_t|z_{tj}=1) = N(\theta_j x_{t-j}, \sigma_j^2), \\ &P(Z_{tj}=1|\cdots) = \pi_j, \qquad j=1\cdots k, \quad t=1,\cdots,n. \end{aligned} \quad (6.10)$$

With this Z, the (conditional) likelihood function is

$$L = \prod_{t=k+1}^{n}\prod_{j=1}^{k}\left[\pi_j\frac{1}{\sqrt{2\pi}\sigma_j}\exp\left\{\frac{1}{2\sigma_j^2}(x_t-\theta_j x_{t-j})^2\right\}\right]^{z_{tj}}. \quad (6.11)$$

We assign the following priors for the parameters as below

$$
\begin{aligned}
&p(k) \propto \frac{\lambda^k}{k!}, \\
&(\pi_1, \cdots, \pi_k), \quad (\pi_1, \cdots, \pi_k) \sim Dirichlet(\lambda), \\
&(\theta_1, \cdots, \theta_k), \quad \theta_j \sim Normal(\mu_0, \sigma_0^2), \quad j = 1, \cdots, k, \\
&(\sigma_1^2, \cdots, \sigma_k^2), \quad \frac{1}{\sigma_j^2} \sim Gamma(\alpha, \beta), \quad j = 1, \cdots, k, \\
&\beta \sim Gamma(h_1, h_2).
\end{aligned}
\tag{6.12}
$$

Since Z is indicator, it can be estimated as in previous chapters. The posterior for z_{tj} is

$$
P(z_{tj} = 1|\cdots) = \frac{\pi_j N(\theta_j x_{t-j}, \sigma_j^2)}{\sum_{j=1}^k \pi_j N(\theta_j x_{t-j}, \sigma_j^2)} = \frac{\pi_j \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left\{\frac{(x_t - \theta_j x_{t-j})^2}{2\sigma_j^2}\right\}}{\sum_{j=1}^k \pi_j \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left\{\frac{(x_t - \theta_j x_{t-j})^2}{2\sigma_j^2}\right\}}. \tag{6.13}
$$

Let C_j denote component j and n_j denote size of component C_j, for $j = 1, \cdots, k$. That is $n_j = \#\{t : x_t \in C_j, t = k+1, \cdots, n\}$.

Next we will calculate the conditional posterior distribution for the weights $(\pi_1, \cdots, \pi_k)$,

$$
\begin{aligned}
f(\pi_1, \cdots, \pi_k|\cdots) \quad &\sim \prod_{t=k+1}^{n} \prod_{j=1}^{k} \pi_j^{Z_{tj}} \pi_1^{\gamma-1} \cdots \pi_{k-1}^{\gamma-1} (1 - \pi_1 - \cdots - \pi_{k-1})^{\gamma-1} \\
&\sim \pi_1^{n_1} \cdots \pi_k^{n_k} \pi_1^{\gamma-1} \cdots \pi_{k-1}^{\gamma-1} (1 - \pi_1 - \cdots - \pi_{k-1})^{\gamma-1} \\
&\sim \pi_1^{\gamma+n_1-1} \cdots \pi_k^{\gamma+n_k-1}.
\end{aligned}
$$

So the posterior for $(\pi_1, \cdots, \pi_k)$ is

$$
f(\pi_1, \cdots, \pi_k|\cdots) = Dirichlet(\gamma + n_1, \cdots, \gamma + n_k), \quad \text{here} \quad n_1 + \cdots + n_k = n - k.
$$

Now calculate the the posterior for $\theta_j|\cdots$

$$\begin{aligned}
f(\theta_j|\cdots) &\sim \prod_{t\in C_j}\frac{1}{\sqrt{2\pi}\sigma_j}\exp\left\{-\frac{(x_t-\theta_j x_{t-j})^2}{2\sigma_j^2}\right\}\times\frac{1}{\sqrt{2\pi}\sigma_0}\exp\left\{-\frac{(\theta_j-\mu_0)^2}{2\sigma_0^2}\right\}\\
&\sim \prod_{t\in C_j}\exp\left\{-\frac{x_{t-j}^2\theta_j^2-2x_{t-j}x_t\theta_j+x_t^2}{2\sigma_j^2}\right\}\times\exp\left\{-\frac{\theta_j^2-2\mu_0\theta_j+\mu_0^2}{2\sigma_0^2}\right\}\\
&\sim \exp\left\{-\frac{\sum\limits_{t\in C_j}x_{t-j}^2\theta_j^2-2\sum\limits_{t\in C_j}x_{t-j}x_t\theta_j}{2\sigma_j^2}-\right\}\times\exp\left\{-\frac{\theta_j^2-2\mu_0\theta_j}{2\sigma_0^2}\right\}\\
&\sim \exp\left\{-\frac{\sigma_0^2\sum\limits_{t\in C_j}x_{t-j}^2\theta_j^2-2\sigma_0^2\sum\limits_{t\in C_j}x_{t-j}x_t\theta_j+\sigma_j^2\theta_j^2-2\mu_0\sigma_j^2\theta_j}{2\sigma_0^2\sigma_j^2}\right\}\\
&\sim \exp\left\{-\frac{\left(\sigma_0^2\sum\limits_{t\in C_j}x_{t-j}^2+\sigma_j^2\right)\theta_j^2-2\left(\sigma_0^2\sum\limits_{t\in C_j}x_{t-j}x_t+\mu_0\sigma_j^2\right)\theta_j}{2\sigma_0^2\sigma_j^2}\right\}\\
&\sim \exp\left\{-\frac{\left(\frac{1}{\sigma_j^2}\sum\limits_{t\in C_j}x_{t-j}^2+\frac{1}{\sigma_0^2}\right)\theta_j^2-2\left(\frac{1}{\sigma_j^2}\sum\limits_{t\in C_j}x_{t-j}x_t+\mu_0\frac{1}{\sigma_0^2}\right)\theta_j}{2}\right\}\\
&\sim \exp\left\{-\frac{\theta_j^2-2\left(\frac{1}{\sigma_j^2}\sum\limits_{t\in C_j}x_{t-j}x_t+\mu_0\frac{1}{\sigma_0^2}\right)\left(\frac{1}{\sigma_j^2}\sum\limits_{t\in C_j}x_{t-j}^2+\frac{1}{\sigma_0^2}\right)^{-1}\theta_j}{2\left(\frac{1}{\sigma_j^2}\sum\limits_{t\in C_j}x_{t-j}^2+\frac{1}{\sigma_0^2}\right)^{-1}}\right\}.
\end{aligned}\tag{6.14}$$

It's easy to see that (6.14) is the kernel of normal density. So

$$\begin{aligned}
f(\theta_j|\cdots)=N\Bigg(&\left(\frac{1}{\sigma_j^2}\sum_{t\in C_j,t\geq k+1}x_{t-j}x_t+\mu_0\frac{1}{\sigma_0^2}\right)\left(\frac{1}{\sigma_j^2}\sum_{t\in C_j,t\geq k+1}x_{t-j}^2+\frac{1}{\sigma_0^2}\right)^{-1},\\
&\left(\frac{1}{\sigma_j^2}\sum_{t\in C_j}x_{t-j}^2+\frac{1}{\sigma_0^2}\right)^{-1}\Bigg).
\end{aligned}$$

Next is to calculate the posterior for $\sigma_j^2, j = 1, \cdots, k$.

$$\begin{aligned}
& f(\sigma_j^2|\cdots) \\
\sim & \prod_{t\in C_j} \frac{1}{\sqrt{2\pi}}(\sigma_j^2)^{-\frac{1}{2}} \exp\left\{-\frac{(x_t-\theta_j x_{t-j})^2}{2\sigma_j^2}\right\} \times \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\sigma_j^2}\right)^{\alpha-1} \exp\left\{-beta\frac{1}{\sigma_j^2}\right\} \\
\sim & \left(\frac{1}{\sigma_j^2}\right)^{\frac{n_j}{2}} \exp\left\{-\sum_{t\in C_j}\frac{(x_t-\theta_j x_{t-j})^2}{2}\frac{1}{\sigma_j^2}\right\} \times \left(\frac{1}{\sigma_j^2}\right)^{\alpha-1} \exp\left\{-\beta\frac{1}{\sigma_j^2}\right\} \\
\sim & \left(\frac{1}{\sigma_j^2}\right)^{\frac{n_j}{2}+\alpha-1} \exp\left\{-\left(\sum_{t\in C_j}\frac{(x_t-\theta_j x_{t-j})^2}{2}+\beta\right)\frac{1}{\sigma_j^2}\right\}. \qquad (6.15)
\end{aligned}$$

(6.15) is the kernel of Gamma density. So the posterior for $\sigma_j^2|\cdots$ is

$$f\left(\frac{1}{\sigma_j^2}|\cdots\right) = \Gamma\left(\frac{n_j}{2}+\alpha, \sum_{t\in C_j}\frac{(x_t-\theta_j x_{t-j})^2}{2}+\beta\right).$$

For the posterior of $\beta|\cdots$,

$$\begin{aligned}
f(\beta|\cdots) &\sim \left(\prod_{j=1}^k \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\sigma_j^2}\right)^{\alpha-1} e^{-\beta\frac{1}{\sigma_j^2}}\right) \times \frac{h_2^{h_1}}{\Gamma(h_1)}\beta^{h_1-1}e^{-h_2\beta} \\
&\sim \beta^{k\alpha}\exp\left\{-\sum_{j=1}^k\frac{1}{\sigma_j^2}\beta\right\}\beta^{h_1-1}e^{-h_2\beta} \\
&\sim \beta^{(k\alpha+h_1)-1}\exp\left\{-\left(\sum_{j=1}^k\frac{1}{\sigma_j^2}+h_2\right)\beta\right\}.
\end{aligned}$$

So

$$f(\beta|\cdots) = \Gamma\left(k\alpha+h_1, \sum_{j=1}^k\frac{1}{\sigma_j^2}+h_2\right). \qquad (6.16)$$

6.7 Simulation Result

Here is the simulation result for the MTD normal example above. The true model has three mixture components. The model is

$$f(x_t|x^{t-1}) = \sum_{j=1}^{3} \pi_j \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left\{ -\frac{(x_t - \theta_j x_{t-j})^2}{2\sigma_j^2} \right\}. \tag{6.17}$$

With the true values of the model given by

True value

j	π	θ	σ
1	.4	1.5	2
2	.3	.9	1
3	.3	.3	.1

We will use both EM algorithm and MCMC Birth-Death process to study this model and compare the simulation results of these two methods. From the simulation we find that when using EM algorithm there are singularities for this model.

6.7.1 EM Algorithm Method

Choosing the initial value of EM algorithm as follows

Initial Value

j	π	θ	σ
1	.45	1.8	2.5
2	.33	1	1.2
3	.22	.8	.9

Repeat the EM algorithm 100 times. Of all 100 repeats there are 5 times the singularities happen. We discard these 5 repeats, using the average of the rest 95 repeats as our estimations. The result is given below(in the parenthesis are bias between true values and the estimates).

Simulation results for EM algorithm

j	π	θ	σ
1	0.3691 (-.0309)	1.4129 (-.0871)	2.0935 (.0935)
2	0.3114 (.0114)	0.8531 (-.0469)	0.9506 (-.0506)
3	0.3195 (.0195)	0.2913 (-.0087)	0.1055 (.0055)

6.7.2 MCMC Birth-Death Process Method

First we want to study whether the result of choosing number of mixtures k is significantly effected or not by choosing different values of λ . Here we set λ to be different values such as $\lambda = 2,\ 5,\ 10,\ 20$. For each λ, we run MCMC 10000 times; Discard the first 5000 times as burning-in, and the rest 5000 is used for our analysis.

Results using Birth-Death Process

λ	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k \geq 6$
2	786	3361	593	189	71
5	703	3209	626	238	224
10	620	3138	605	355	282
20	589	3156	933	186	136

From the table above, we see that for different value of λ, the simulation results of k are the same: $k = 3$ appears much more often than any other values. From this simulation we can conclude that different values of λ are not significant for Birth-Death process method.

Next is to estimate parameters in the model. We will run the Birth-Death process for 6500 times. Throwing away the first 2500 as burning-in, we use the rest 4000 to get our estimation. Of these 4000 jumps, there are 2289 times $k = 3$. Based on this, we can get the estimation of the parameters using the sample mean (in the parenthesis are bias).

Simulation Results for MCMC Method

j	π	θ	σ
1	0.4040 (.0040)	1.5343 (.0343)	2.0435 (.0435)
2	0.3073 (.0073)	0.9139 (.0139)	0.9382 (-.0618)
3	0.2887 (-.0113)	0.2940 (-.0060)	0.1091 (.0091)

From this comparison, we can see superiorities of MCMC methods. First, EM algorithm here has singularity problem but MCMC does not. Second, for all 9 parameters estimations, 7 estimations from MCMC are more accurate than that from EM algorithm (there are 7 bias with MCMC method smaller than the bias with EM algorithm method). Thirdly, for MCMC, estimation of the parameters is from the same data, so we can construct the Confidence Intervals for all the parameters.

Based on the results of MCMC, we can construct the 95% confidence interval for the parameters.

95% Confidence Interval

$\pi_1 \in (0.348, 0.458)$	$\theta_1 \in (1.457, 1.612)$	$\sigma_1 \in (1.593, 2.692)$
$\pi_2 \in (0.252, 0.363)$	$\theta_2 \in (0.892, 0.935)$	$\sigma_2 \in (0.551, 1.151)$
$\pi_3 \in (0.259, 0.317)$	$\theta_3 \in (0.287, 0.295)$	$\sigma_3 \in (0.090, 0.132)$

From above of the 95% confidence intervals, we can see for almost all parameters except θ_3, the 95% confidence intervals will cover the true value. The confidence intervals help to verify the excellent performance of Birth-Death process method.

6.7.3 Summary

MCMC method is another way of doing parameter estimation and model selection for the MTD/BMTD models. Different from EM method of finding the point estimations that (locally) maximize the posterior (likelihood), MCMC method samples the value of the parameters $(k, \boldsymbol{\pi}, \boldsymbol{\phi})$ from the posterior distribution. After getting the samples, model selection of k is given by the mode of the samples of k. After we set up k, other parameters $(\boldsymbol{\pi}, \boldsymbol{\phi})$ are estimated by the average of the samples having the corresponding k.

How to choose a suitable prior? Sometimes we may have strong prior information about the parameters in the model. For example, if we have a training data, we can get vague knowledge about how many mixture components are there in the model, and we can get a more accurate value of the parameters in the prior distribution.

For other times we may have little information about the parameters and the number of components in the MTD model. Under this condition people (Rubin, Andrew 2003) usually suggest non-informative prior distributions which present ignorance or lack of information about the parameters in the model. An example is Jeffery prior. In our example, we select non-informative prior for the parameters which have nothing to do with singularities. Another widely used method is to use the conjugate priors for the parameters. With conjugate priors, the posterior is easy to be calculated. In our example, we choose conjugate priors for σ_j^2. It brings us two facilities: first it's easy to get the posterior; second it guarantees that the posterior will never be infinity for any value of the parameters.

Chapter 7

Dirichlet Process and Its Application in Mixture Model

In Chapter 3 and Chapter 4 we pre-fix number of mixtures k and assign priors for other parameters in the model, then we use EM algorithm to maximize the posterior. Compared to non-Bayesian method, the proposed Bayesian method can solve the problem of singularity and the estimators are consistent as proved in Chapter 5. In Chapter 6 we treat k as random and also assign a discrete prior for k. Then we use MCMC to sample from the posterior. k is estimated by the mode from its samples.

However, in reality when we get the data, we don't know whether the data comes from a mixture model or not; or if it comes from a mixture model, we don't know how many mixtures there are in the model. Rather than being explicitly expressed in the mixture model, each observation within the data sets can be associated with a (set of) parameter. We assume the observations with the same parameters coming from the same mixture component. Because most of the time the parameters are continuous on their support, the probability of two parameters being equal is zero. To solve this problem, Neal (2000) introduced a discrete type of prior for the parameters. Then for the posterior, there is a positive probability that two parameters are equal to each other. This discrete type of prior we used here is called Dirichlet process(Ferguson, 1973; Neal 2000). In this chapter we will first give an introduction of Dirichlet process mixtures. Next we will introduce its background and explain in detail how to use Dirichlet process prior in our mixture models.

7.1 Dirichlet Process and Chinese Restaurant Method

Modeling a distribution as a mixture of simpler distributions is useful both as a nonparametric density estimation method and as a way of identifying latent classes that can explain the dependencies observed between variables. Mixtures with a countably infinite number of components can reasonably be handled in a Bayesian framework by employing a prior distribution for mixing proportions, such as a Dirichlet process. Dirichlet process mixture models has become computationally feasible with the development of Markov chain methods for sampling from the posterior distribution of the parameters of the component distributions. Methods based on Gibbs sampling can easily be implemented for models based on conjugate prior distributions.

Bayesian nonparametric studies were initially introduced by Ferguson (1973) when he introduced Dirichlet processes for modeling random distributions. With a Dirichlet process prior, Blackwell and MacQueen (1973) showed that the marginal distribution of the missing values or latent variables had a Polya structure. This result had inspired many researchers to work on computational procedures for Bayesian nonparametric methods. Lo et al. (1996) introduced a sampling method for sampling partitions. Instead of generating missing values, they evaluated the posterior expectation by sampling partitions that are sufficient statistics of the missing values. Y.W. Teh, M.I. Jordan, M.J. Beal and D.M. Blei (2006) introduced the hierarchical Dirichlet process for mixture models and cluster analysis. This model has more flexibility in modeling data with mixture characters.

Next we will brifly introduce how Dirichlet process is used in the mixture models.

7.2 Dirichlet Distribution

The Dirichlet Distribution is a multi-parameter extension of Beta Distribution. It defines a distribution over distributions on the discrete probability space. Let $\boldsymbol{\pi} = \{\pi_1, \pi_2, \cdots, \pi_n\}$ be a probability distribution on the discrete space $\mathscr{X} = \{\mathscr{X}_1, \mathscr{X}_2, \cdots, \mathscr{X}_n\}$ such that $P(X = \mathscr{X}_i) = \pi_i$. The Dirichlet distribution over $\boldsymbol{\pi}$ is given by

$$P(\boldsymbol{\pi}|\alpha, H) = \frac{\Gamma(\alpha)}{\prod_{i=1}^{n} \Gamma(\alpha h_i)} \prod_{i=1}^{n} \pi_i^{\alpha h_i - 1}$$

where $H = \{h_1, h_2, \cdots, h_n;\ h_i > 0\}$ is the *base measure* defined on $\mathscr{X}$ and is also the mean value of $\boldsymbol{\pi}$. α is a *strength parameter* that says how concentrated the distribution is around H. Both $\boldsymbol{\pi}$ and H are discrete probability distributions and their sum is to unity.

7.3 Dirichlet Process (DP)

For a random distribution G to be distributed according to a DP, its marginal distribution has to be Dirichlet distributed. Specifically, let H be a distribution over Θ and α be a positive real number. Then for any finite measurable partition $A_1, \cdots, A_r$ of Θ, the vector $(G(A_1), \cdots, G(A_r))$ is random since G is random.

Ferguson (1973) gave the definition of Dirichlet process as below: A random distribution G is Dirichlet process distributed with base distribution H and concentration parameter α, written as $G \sim DP(\alpha, H)$ if

$$(G(A_1), \cdots, G(A_r)) \sim Dir(\alpha H(A_1), \cdots, \alpha H(A_r))$$

for every finite measurable partition $A_1, \cdots, A_r$ of Θ.

Ferguson also showed that the expectation and variance of G are

$$\begin{aligned} E(G(A)) &= H(A), \\ V(G(A)) &= H(A)(1 - H(A))/(1 + \alpha). \end{aligned}$$

So, as $\alpha \to +\infty$, G looks more like H.

7.4 Posterior Distribution

Let Θ be a space and A is the subspace of Θ, that is $A \subset \Theta$. Let

$$\begin{aligned} &\theta|G \sim G(\theta), \qquad i.e.\ \ P(\theta \in A \subset \Theta|G) = G(A), \\ &G \sim DP(\alpha, H). \end{aligned} \tag{7.1}$$

Since G is a random distribution, we can draw samples from G.

Let $\theta_1, \cdots, \theta_n$ be a sequence of independent draws from G. Note that these $\theta_i's$ take values in Θ since G is a distribution over Θ. We want to calculate the posterior of $G|\theta_1, \cdots, \theta_n$.

Let $A_1, \cdots, A_r$ be a finite partition of Θ, and $n_k = \#\{i : \theta_i \in A_k\}$. After some calculations (See Appendix in this chapter), we have

$$(G(A_1), \cdots, G(A_r))|\theta_1, \cdots, \theta_n \sim Dir(\alpha H(A_1) + n_1, \cdots, \alpha H(A_r) + n_r).$$

Since it's true for all finite partition, from definition of DP, we know $G|\theta_1, \cdots, \theta_n$ is a DP.

After calculations (See calculation in appendix), we get

$$G|\theta_1, \cdots, \theta_n \sim DP\left(\alpha + n, \frac{\alpha H}{\alpha + n} + \frac{\sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}\right). \tag{7.2}$$

Note 1. *Priors weight base on α, and empirical distribution has weight on n. i.e., as the number of observations grows and $n >> \alpha$, the posterior is dominated by empirical distribution.*

7.5 Representation of Dirichlet Process

7.5.1 Blackwell-MacQueen Urn Representation

Let $\theta_1, \theta_2, \cdots$ be samples from distribution G. Consider the conditional distribution of $\theta_{n+1}|\theta_1, \cdots, \theta_n$ with G integrated out. We know $\theta_{n+1}|G, \theta_1, \cdots, \theta_n \sim G$ for $A \subset \Theta$.

$$\begin{aligned} P(\theta_{n+1} \in A|\theta_1, \cdots, \theta_n) &= E(G(A)|\theta_1, \cdots, \theta_n) \\ &= \frac{\alpha H(A)}{\alpha + n} + \frac{\sum_{i=1}^n \delta_{\theta_i}(A)}{\alpha + n}. \end{aligned} \tag{7.3}$$

Thus, $\theta_{n+1}|\theta_1, \cdots, \theta_n \sim \frac{\alpha H}{\alpha+n} + \frac{\sum_{i=1}^n \delta_{\theta_i}}{\alpha+n}$, which is the posterior base distribution given $\theta_1, \cdots, \theta_n$.

This model can be explained by the Polya Urn scheme.

There are α balls in an urn, of which $\alpha H(A_1)$ in 1^{st} color, and $\alpha H(A_2)$ in 2^{nd} color, $\cdots$, $\alpha H(A_r)$ in r^{th} color. Now draw a ball from the urn and replace it by two balls of the same color as the one drawn.

So,

$$\begin{gathered} P(\theta_1 \in A_i) = \frac{\alpha H(A_i)}{\sum_{i=1}^r \alpha H_{(A_i)}} = \frac{\alpha H(A_i)}{\alpha} = H(A_i), \\ P(\theta_2 \in A_i|\theta_1) = \frac{\alpha H(A_i) + \delta(\theta_1 \in A_i)}{\alpha + 1} = \frac{\alpha H(A_i) + \delta_{\theta_1}(A_i)}{\alpha + 1}, \\ \vdots \\ P(\theta_{n+1} \in A_i|\theta_1, \cdots, \theta_n) = \frac{\alpha H(A_i)}{\alpha + n} + \frac{\sum_{j=1}^n \delta_{\theta_j}(A_i)}{\alpha + n}. \end{gathered}$$

So

$$\theta_{n+1}|\theta_1, \cdots, \theta_n \sim \frac{\alpha H}{\alpha + n} + \frac{n}{\alpha + n} \frac{\sum_{i=1}^n \delta_{\theta_i}}{n}. \tag{7.4}$$

That is, θ_{n+1} with probability proportional to $\frac{\alpha}{\alpha+n}$ comes from H, and with probability proportional to $\frac{n}{\alpha+n}$ equals to one of existing $\theta_i,\ i=1$ to n. This is Blackwell-MacQueen representation of Dirichlet process.

In section 7.6 we will show how to use this representation to study mixture models.

7.5.2 Chinese Restaurant Representation

From the previous representation of Dirichlet process, we can see given θ_1, the new sample θ_2 may equal to θ_1 or it may be a new value different from θ_1; Given θ_1 θ_2, a new sample θ_3 may equal to θ_1, θ_2 or a new different value. So, if we continue to sample $\theta_1, \cdots, \theta_n$ in this way, some of them will take the same value.

Denote the unique values among $\theta_1, \cdots, \theta_n$ as $\theta_1^*, \cdots, \theta_m^*$; n_k is the number of repeats of θ_k^*; that is $n_k = \#\{i : \theta_i = \theta_k^*\}$, $k = 1$ to m; The predictive distribution (7.4) can be rewritten as

$$\theta_{n+1}|\theta_1, \cdots, \theta_n \sim \frac{\alpha H}{\alpha+n} + \frac{\sum_{k=1}^m n_k \delta_{\theta_k^*}}{\alpha+n} \tag{7.5}$$

From (7.5) we can see the unique value of $\theta_1, \cdots, \theta_n$ induces a partition $\{T_1, \cdots, T_m\}$ of these $\theta's$ with T_k=$\{\theta_i : \theta_i = \theta_k^*,\ i = 1$ to $n\}$; On each T_k , $\theta_i's$ all take the same value θ_k^*.

- We assign labels for $\{\theta_1, \theta_2, \cdots, \theta_n\}$ as $\{c_1, c_2, \cdots, c_n\}$;
- If $c_i = j$ means $\theta_i = \theta_j^*$, $i = 1, \cdots, n$, $j = 1, \cdots, m$, and we say customer i (θ_i and the corresponding x_i) is on table T_j;

- The sampling above is equivalent to

  - First sample these $\{c_1, c_2, \cdots, c_n\}$ to get the partition $\{T_1, \cdots, T_m\}$;

    $$\begin{cases} P(c_{i+1} = c | c_1, \cdots, c_i) \propto \dfrac{n_c}{i+\alpha}, \quad c \in \{c_1, c_2, \cdots, c_i\}, \\ P(c_{i+1} \neq c_j, \text{ for all } j = 1, \cdots, i | c_1, \cdots, c_i) \propto \dfrac{\alpha}{i+\alpha}, \end{cases}$$

    with $n_c = \#\{j : c_j = c, \quad j = 1, \cdots, i\}$

  - Then sample the parameters θ_j^* on table T_j from the distribution H.

So we show how to use these Dirichlet representations to study mixture model.

7.6 Mixture Model Expressed via Dirichlet Process

Neal (2000) gave the brief introduction of Dirichlet process in the Mixture Normal distributions.

The mixture model can be expressed in Dirichlet process model

$$\begin{aligned} &x_i|\theta_i \sim F(\theta_i) \\ &\theta_i|G \sim G \\ &G \sim DP(\alpha, H) \end{aligned} \tag{7.6}$$

Let θ_{-i} denote all the $\theta_j's$ for $j \neq i$ and $x_1, \cdots, x_n$ be the data.

From previous we know the prior for θ_i can be obtained from

$$P(\theta_i|\theta_{-i}) = \frac{1}{n-1+\alpha} \sum_{j \neq i} \delta(\theta_j) + \frac{\alpha}{n-1+\alpha} H(\theta_i).$$

Combined with the likelihood, it yields the following conditional distribution for use in Gibbs sampling

$$\theta_i|\theta_{-i}, x_i \sim b \cdot \alpha \cdot H(\theta_i)F(x_i;\theta_i) + b\sum_{j\neq i}^{n} \delta(\theta_j)F(x_i;\theta_j), \tag{7.7}$$

here

$$\begin{aligned} b &= \left[\alpha \cdot q_0(x_i) + \sum_{j=1;j\neq i}^{n} F(x_i;\theta_j)\right]^{-1}, \\ q_0(x_i) &= \int_\theta F(x_i;\theta)dH(\theta). \end{aligned} \tag{7.8}$$

It can be observed that $q_0(x_i)$ is actually the marginal distribution of x_i. Let's denote $h(\theta_i|x_i) = \frac{H(\theta_i)F(x_i;\theta_i)}{\int_\theta F(x_i;\theta)dH(\theta)}$ and (7.7) can be rewritten as

$$\theta_i|\theta_{-i}, x_i \sim b \cdot \alpha \cdot q_0(x_i) \cdot h(\theta_i|x_i) + b\sum_{j\neq i} \delta(\theta_j)F(x_i;\theta_j). \tag{7.9}$$

This can be written in a form that demonstrates the mixture nature of the marginal posterior on θ_i and also give a simple algorithm for sampling from $\theta_i|\theta_{-i}, x_i$

$$\theta_i|\theta_{-i}, x_i \begin{cases} = \theta_j, & \text{with probability } \; b \cdot F(x_i;\theta_j), \\ \sim h(\theta|x_i) & \text{with probability } \; b \cdot \alpha \cdot q_0(x_i). \end{cases} \tag{7.10}$$

A Gibbs sampling algorithm using (7.10) can be designed to perform sampling on the space of θs.

In a conjugate model, the distributions F and H are being conjugated and the integration in the calculation of q_0 can be performed explicitly.

7.6.1 A Normal Mixture Example

A normal mixture is used as an example. As an example of (7.6), we take F as normal ddistribution $N(\mu_i, 1)$. Then (7.6) is rewritten as

$$\begin{aligned} &x_i|\mu_i \sim N(\mu_i, 1) \\ &\mu_i \sim G(\mu) \\ &G \sim DP(\alpha, H) \\ &H \sim N(0, 1) \end{aligned} \tag{7.11}$$

Using the formulas in (7.8) we get

$$\begin{aligned} q_0(x_i) &= \frac{1}{2\sqrt{\pi}} \exp\left\{\frac{-x_i^2}{4}\right\}, \\ h(\mu|x_i) &= \frac{1}{\sqrt{2\pi}} \exp\left\{-(\mu - x_i/2)^2\right\}. \end{aligned} \tag{7.12}$$

So the Gibbs sampler becomes

$$\mu_i|\mu_{-i}, x_i \begin{cases} = \mu_j, & \text{with probability proportional to } F(x_i; \mu_j), \\ \sim h(\mu|x_i), & \text{with probability proportional to } \alpha q_0(x_i). \end{cases} \tag{7.13}$$

That is, $\mu_i|\mu_{-i}, x_i$ equals to one of the existed $\mu_j's$ with probability proportional to $F(x_i; \mu_j)$. $\mu_i|\mu_{-i}, x_i$ equals to a new value sampled from density $h(\mu|x_i)$ with probability proportional to $\alpha q_0(x_i)$.

Algorithm 1. *Initializing the value of $\boldsymbol{\mu}^{(0)}$, we can sample the $\boldsymbol{\mu}^{(j)}$ in the following*

way

$$\begin{aligned}
&\textit{Sample}\ \ \mu_1^{(j)}\ \ \textit{from}\ \ \mu_1|\mu_2=\mu_2^{(j-1)},\ \mu_3=\mu_3^{(j-1)},\ \ \cdots,\ \mu_n=\mu_n^{(j-1)},\\
&\textit{Sample}\ \ \mu_2^{(j)}\ \ \textit{from}\ \ \mu_2|\mu_1=\mu_2^{(j)},\ \ \mu_3=\mu_3^{(j-1)},\ \ \cdots,\ \mu_n=\mu_n^{(j-1)},\\
&\qquad\quad\vdots\\
&\textit{Sample}\ \ \mu_n^{(j)}\ \ \textit{from}\ \ \mu_n|\mu_1=\mu_1^{(j)},\ \ \mu_2=\mu_2^{(j)},\ \ \cdots,\ \mu_{n-1}=\mu_{n-1}^{(j)}.
\end{aligned}$$

After we get $\boldsymbol{\mu}^{(1)},\cdots,\boldsymbol{\mu}^{(M)},\boldsymbol{\mu}^{(M+1)},\cdots,\boldsymbol{\mu}^{(M+N)}$. The first M samples will be dropped as burning-in. For the rest N samples, denote the number of unique values in each $\boldsymbol{\mu}_i, i=1\cdots N$ as $k_1,\cdots,k_N$. Number of mixtures is estimated by mode of $k_1,\cdots,k_N$, denoted as k_0. Select those $\boldsymbol{\mu}_i, i=1\cdots N$ whose $k_i=k_0$ and their mean is the estimation of μ.

7.7 Mixture Model with DP Chinese Restaurant Representation

Algorithm 1 is easy and straightforward to understand. One problem with Algorithm 1 is that the convergence rate is slow. The problem is that there are often groups of observations with high probability that are associated with the same θ. But this algorithm cannot change the θ for more than one observation simultaneously.

If we use Chinese Restaurant process representation of Dirichlet process, it can help to speed up the convergence speed. we can rewrite the model as

$$\begin{cases} x_i|(\boldsymbol{\theta}, c_i) \sim F(\theta_{c_i}) \\ \\ \theta_{c_i} \sim H \\ \begin{cases} P(c_i = c_j|c_{-i}) \propto \dfrac{n_{j,-i}}{n-1+\alpha}, \quad c_j \in c_{-i} = \{c_s, s \neq i, \ s = 1 \text{ to } n\} \\ \\ P(c_i \neq c_j, \text{ for all } j \neq i|c_{-i}) \propto \dfrac{\alpha}{n-1+\alpha}, \end{cases} \end{cases}$$

where $n_{j,-i} = \#\{ \text{ of } c_s = c_j, \ s \neq i, \ s = 1 \text{ to } n\}$.

These $\{c_1, \cdots, c_n\}$ give a partition of $\{x_1, \cdots, x_n\}$. The partition is denoted as $\mathbb{P} = \{T_1, \cdots, T_p\}$, which looks like p tables; Observations on the same table have the same parameter θ; Different partition $\mathbb{P}$ gives different mixture model.

The posterior of c_i given $x_i, \ c_j \ (j \neq i), \ i = 1$ to n, is

$$\begin{cases} P(c_i = c_j|c_{-i}, x_i, \theta) = b\frac{n_{j,-i}}{n-1+\alpha} F(x_i, \theta_{c_j}), \quad j \neq i \\ \\ \\ P(c_i \neq c_j, \text{ for all } j \neq i|c_{-i}, x_i, \theta) = b\frac{\alpha}{n-1+\alpha} \int F(x_i, \theta)H(\theta)d\theta, \end{cases} \tag{7.14}$$

where

$$b = \left(\alpha q_0(x_i) + \sum_{j=1; j\neq i}^{n} F(x_i; \theta_j) \right)^{-1},$$

$$q_0(x_i) = \int_\theta F(x_i; \theta)H(\theta)d\theta.$$

Algorithm 2. *Assume the current state of Markov Chain consist of* $(c_1, \cdots, c_n)$ *and* $\vec{\theta} = (\theta_c : c \in \{c_1, \cdots, c_n\})$. *The current partition is* $\mathbb{P}^{(s)}$. *Repeatedly sample as follows to get the new partition* $\mathbb{P}^{(s+1)}$*:*

1. *For* $i = 1, \cdots, n$*: if the present table that* c_i *is on has no other observation, that is,* $n_{-i,c_i} = 0$*, remove* θ_{c_i} *from the state. Draw a new value for* c_i *from* $c_i|c_{-i}, x_i, \vec{\theta}$

as defined by equation (7.14). If the new c_i is not on the existing table, we get a new table; Otherwise, if the new c_i is on existing table, we update the cardinality of that table.

If the new c_i is not associated with any observation, draw a value for θ_{c_i} from the posterior:

$$P(\theta|x_i) \propto F(x_i, \theta) * H(\theta).$$

2. *Now suppose we got the updated value of $(c_1, \cdots, c_n)$. Denote the partition as $\mathbb{P}^{(s+1)} = \{T_1, \cdots, T_p\}$;*

3. *For each $T_j \in \mathbb{P}^{(s+1)}$: Draw a new value for $\theta^{(s+1)}$ on T_j from the posterior distribution based on the prior H and all the data points on T_j, that is:*

$$P(\theta|\mathbf{y}) \propto \prod_{i \in T_j} F(x_i, \theta) * H(\theta).$$

After we get the samples, the parameters in the model is estimated in the following way.

7.7.1 Determine the Number of Components

We will let the MCMC run $M + N$ times, and regard the first M times as a burning in. The rest of the data points from $M + 1$ to $M + N$ will be left for our study.

In the Dirichlet process, for each iteration, we get the partition $\mathbb{P}^{(i)}$, $i \in \{M + 1, \cdots, M + N\}$. Each partition $\mathbb{P}^{(i)}$ gives the number of tables $n(\mathbb{P}^{(i)})$, which is essentially the number of components in iteration i. It's natural to determine the number of components as the posterior mode of $n(\mathbb{P})$. With the MCMC iterates of $n(\mathbb{P}^{(i)})$, we can estimate the number of components or the posterior mode by selecting k which maximizes $\frac{1}{N}\sum_{i=M+1}^{M+N} I(n(\mathbb{P}^{(i)}) = k)$, the sample proportion of $n(\mathbb{P}) = k$.

7.7.2 Parameter Estimation

After getting the number of components $k = k_0$. A set that collects all indices of partition with $n(\mathbb{P}^{(j)}) = k_0$ is defined as $\mathcal{N} = \{j : n(\mathbb{P}^{(j)}) = k_0\}$. Parameter estimation could be defined as the average of the parameters over the set $\mathcal{N}$. That is, $\theta = \frac{1}{N_0}\sum_{j\in\mathcal{N}} \theta^j$ where N_0 is the size of $\mathcal{N}$.

7.7.3 A Simple Example to Explain Algorithm 2

Here a simple explanation of the second algorithm is shown below.

Suppose we have data $x_1, x_2, \cdots, x_{10}$; Current status the label c_i and corresponding θ_{c_i} is

Table1: data, tables, and parameters

x_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
c_i	2	3	2	1	2	3	2	3	3	2
parm	θ_2	θ_3	θ_2	θ_1	θ_2	θ_3	θ_2	θ_3	θ_3	θ_2

From the above table, we can write down the current partition of subscript as: $\mathbb{P}^t = \{T_1 = \{1, 3, 5, 7, 10 \mid \theta_2\},\ T_2 = \{2, 6, 8, 9 \mid \theta_3\},\ T_3 = \{4 \mid \theta_1\}\}$.

With Algorithm 2, a detailed explanation of transfer from current status $\mathbb{P}^t$ to next status $\mathbb{P}^{t+1}$ is given as below:

- Remove c_1 and get new partition $\mathbb{P}_{-1} = \{T_1 = \{3, 5, 7, 10 \mid \theta_2\},\ T_2 = \{2, 6, 8, 9 \mid \theta_3\},\ T_3 =$

$\{4 \mid \theta_1\}\}$;

- Reseat c_1 by the given probability; Suppose c_1 is still on T_1; Now $\mathbb{P} = \{T_1 = \{1,3,5,7,10 \mid \theta_2\},\ T_2 = \{2,6,8,9 \mid \theta_3\},\ T_3 = \{4 \mid \theta_1\}\}$;
- Remove c_2 and get new partition $\mathbb{P}_{-2} = \{\{1,3,5,7,10 \mid \theta_2\}, \{6,8,9 \mid \theta_3\}, \{4 \mid \theta_1\}\}$;
- Reseat c_2 by the given probability; Suppose 2 is on new table rather than existing table; now $\mathbb{P} = \{T_1 = \{1,3,5,7,10 \mid \theta_2\},\ T_2 = \{6,8,9 \mid \theta_3\},\ T_3 = \{4 \mid \theta_1\},\ T_4 = \{2 \mid \theta_4\}\}$ by sampling θ_4 from $F(x_2, \theta) * H(\theta)$;
- Remove c_3 and do the same thing; Suppose it's on T_1; now $\mathbb{P} = \{T_1 = \{1,3,5,7,10 \mid \theta_2\},\ T_2 = \{6,8,9 \mid \theta_3\},\ T_3 = \{4 \mid \theta_1\},\ T_4 = \{2 \mid \theta_4\}\}$;
- Remove c_4; Since T_4 has only 4; We remove the corresponding $\theta = \theta_1$; $\mathbb{P}_{-4} = \{T_1 = \{1,3,5,7,10 \mid \theta_2\},\ T_2 = \{6,8,9 \mid \theta_3\},\ T_4 = \{2 \mid \theta_4\}\}$;
- Reseat c_4. Suppose it's on T_4, now we have $\mathbb{P} = \{T_1 = \{1,3,5,7,10 \mid \theta_2\},\ T_2 = \{6,8,9 \mid \theta_3\},\ T_4 = \{2,\ 4 \mid \theta_4\}\}$;
- For $c_5, \ \cdots, c_{10}$ we do the same thing. Then we get the updated partition $\mathbb{P}^{t+1}$ is

$$\mathbb{P}^{t+1} = \{T_1 = \{1,3,5,7\},\ T_2 = \{6,8,9,10\},\ T_3 = \{2,4\}\};$$

- Updated parameter θ on each $T_j \in \mathbb{P}^{t+1}$ is

$$P(\theta|\mathbf{x}) \propto \prod_{i \in T_j} F(x_i, \theta) * H(\theta).$$

The updated partition $\mathbb{P}^{t+1}$ is:

Table2: Updated

x_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
c_i	2	1	2	1	2	3	2	3	3	3
parm	θ_2	θ_1	θ_2	θ_1	θ_2	θ_3	θ_2	θ_3	θ_3	θ_3

7.8 Simulation Result

7.8.1 An Easy Example

There are many examples of simulation for Dirichlet process approach in normal mixture models (see Neal (2000)). Here we show a simple example of normal mixture where the variance is set to 1 and the number of mixture component is set to 2. The more complicated examples are shown in Chapter 8. The normal mixture density is

$$f(x_t) = \sum_{j=1}^{k} \pi_j \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_t-\theta_j)^2}{2}}.$$

For the simulation, we will choose $k = 3$, that is, there are three components of the normal mixture model. The true value is given by

True Parameters

π_1	π_2	π_3	θ_1	θ_2	θ_3
.3	.45	.25	1.2	2.5	-2.4

The simulation result is (sample size $n = 200$)

Estimation

π_1	π_2	π_3	θ_1	θ_2	θ_3
.3016	.4483	.2501	1.1906	2.4669	-2.4010

From the simulation it is seen that Dirichlet process approach works well in the mixture models. Next we will show some comparison of Dirichlet process approach with EM algorithm and different sample size n.

7.8.2 Parameter Estimation and Comparison

In the second simulation, we will let $k = 2, n = 60$ (here use $k = 2$ as an example, in next chapter we use $k = 4$ as an example to compare). The true values of parameters are

True Value of parameters

parameter	π_1	π_2	θ_1	θ_2
estimation	.3	.7	1	-1

Now we will use EM algorithm. Repeat the simulation 20 times for both methods. The simulation result of the average of 20 repeats is shown below

Simulation Results for EM algorithm

parameter	π_1	π_2	θ_1	θ_2
estimation	.2939	.7061	0.9704	-1.0816

Using the same model for Dirichlet process mixture models, using Algorithm 2, the result is

Simulation for DP, n=60 rep=20

parameter	π_1	π_2	θ_1	θ_2
estimation	.2995	.7005	0.9953	-0.9434

Compared with these two methods, we find both methods yield very good estimations for the parameters and that the difference of parameter estimation between these two methods is not great. That's probably because the true model is pretty well organized.

For the same model, using Dirichlet process mixture models, the sample size increased from 60 to 200, and the replication increased to 100, the result is

Simulation for DP, n=200 rep=100

parameter	π_1	π_2	θ_1	θ_2
estimation	.2996	.7004	1.0071	-0.9721

It looks like the improvement is not significant when we increase the sample size and the iterations because the estimation from the small sample size is already good enough.

Here in the simulation we use algorithm 2 rather than algorithm 1 because algorithm 2 will converge faster than algorithm 1 (Neal 2000).

7.9 Proofs

Calculation of posterior of Dirichlet Process G

First let's have a look at the conjugate relation between Dirichlet Distribution and Multinomial distribution.

$$(\pi_1, \pi_2, \cdots, \pi_r) \sim Dir(\alpha_1, \alpha_2, \cdots, \alpha_r),$$

$$z|(\pi_1, \pi_2, \cdots, \pi_r) \sim Discrete(\pi_1, \pi_2, \cdots, \pi_r).$$

Where z is a multinomial random variable, taking values on $i \in \{1, 2, \cdots, r\}$ with probability π_i. Then we get

$$P(z = j) = E(I(z = j)) = E(\pi_j) = \frac{\alpha_j}{\sum\limits_{i=1}^{r} \alpha_i} \tag{7.15}$$

and

$$\begin{aligned}
(\pi_1, \pi_2, \cdots, \pi_r)|z = j &= \frac{f(\pi_1, \pi_2, \cdots, \pi_r) \times P(z = j|(\pi_1, \pi_2, \cdots, \pi_r))}{P(z = j)} \\
&= \frac{\frac{\Gamma\left(\sum\limits_{i=1}^{r} \alpha_i\right)}{\prod\limits_{i=1}^{r} \Gamma(\alpha_i)} \prod_{i=1}^{r} \pi_i^{\alpha_i - 1} \times \pi_j}{\frac{\alpha_j}{\sum\limits_{i=1}^{r} \alpha_i}} \\
&= \frac{\Gamma\left(\sum\limits_{i=1}^{r} (\alpha_i + \delta_j(z))\right)}{\prod\limits_{i=1}^{r} \Gamma(\alpha_i + \delta_j(z))} \prod_{i=1}^{r} \pi_i^{\alpha_i + \delta_j(z) - 1}.
\end{aligned}$$

So

$$\begin{aligned}
&P(z = j) = \frac{\alpha_j}{\sum\limits_{i=1}^{r} \alpha_i}, \\
&(\pi_1, \pi_2, \cdots, \pi_r)|z = j \sim Dir(\alpha_1 + \delta_j(z), \cdots, \alpha_r + \delta_j(z)).
\end{aligned} \tag{7.16}$$

Now let's turn to Dirichlet process.

$$\begin{aligned}
&(G(A_1), \cdots, G(A_r)) \sim Dir(\alpha H(A_1), \cdots, \alpha H(A_r)), \\
&P(\theta \in A_i|G) = G(A_i).
\end{aligned}$$

From above, we have:

$$\begin{aligned}
&P(\theta \in A_i) = H(A_i), \\
&(G(A_1), \cdots, G(A_r))|\theta \sim Dir(\alpha H(A_1) + \delta_\theta(A_1), \cdots, \alpha H(A_r) + \delta_\theta(A_r)).
\end{aligned} \tag{7.17}$$

Since (7.17) is correct for any partition $(A_1, A_2, \cdots)$, so we have that the posterior of $G|\theta$ is Dirichlet process $DP\left(\alpha + 1, \frac{\alpha H + \delta_\theta}{\alpha + 1}\right)$.

Generally, we have:

$$G|\theta_1, \cdots, \theta_n \sim DP\left(\alpha + n, \frac{\alpha}{\alpha + n} H + \frac{n}{\alpha + n} \frac{\sum_{i=1}^{n} \delta_{\theta_i}}{n}\right). \tag{7.18}$$

Chapter 8

BMTD Models and Its Dirichlet Process Expression

In this section, we will introduce a new method called Dirichlet process (DP) mixtures to study the BMTD model based on the mixture of bivariate distributions discussed before. To make the posterior easier to calculate, we will assign conjugate priors for the parameters. In order to get the conjugate priors, we need to add some restrictions for the re-parameterizations of the original parameters. For the general model (4.2), we will prove that the following general parameterization of λ and β given below will satisfy our requirements

$$\begin{aligned}\lambda &= \theta_1 \cdot g_1\left(x^{t-1}, y^{t-1}\right),\\ \beta &= (\theta_2)^{1/\phi} \cdot g_2\left(x^{t-1}, y^{t-1}\right)\end{aligned} \tag{8.1}$$

where g_1 and g_2 are general functions, and $(x^{t-1}, y^{t-1}) = (x_{t-1}, \cdots, x_1; y_{t-1}, \cdots, y_1)$.

From (4.2), the corresponding BMTD model is

$$f(x_t, y_t|x^{t-1}, y^{t-1}) = \sum_{j=1}^{k} \pi_j C_j x_t^{\delta+\gamma+1/\phi-1} y_t^{\delta} e^{-x_t^{\alpha}(\lambda_j + y_t^{\phi}/\beta_j^{\phi})}, \quad x_t > 0, \; y_t > 0. \qquad (8.2)$$

Re-parametrizations of λ_j and β_j are

$$\begin{aligned} \lambda_j &= \theta_{j,1} \cdot g_1\left(x^{t-1}, y^{t-1}\right), \\ \beta_j &= (\theta_{j,2})^{1/\phi} \cdot g_2\left(x^{t-1}, y^{t-1}\right). \end{aligned} \qquad (8.3)$$

This re-parameterizations will cover the examples by Hassan and Lii (2006). For example, if let $\theta_{j,1} = \theta_j$, $\theta_{j,2} = 1/\beta_j$, $g_1\left(x^{t-1}, y^{t-1}\right) = \frac{1}{x_{t-j}e^{-y_{t-j}}}$ and $g_2\left(x^{t-1}, y^{t-1}\right) == 1$, we will get $\lambda_j = \frac{\theta_j}{x_{t-j}e^{-y_{t-j}}}$ and $\beta_j = \frac{1}{\beta_j}$, that is the first example in their paper; If we set $\theta_{j,1} = \delta_j$ and $g_1\left(x^{t-1}, y^{t-1}\right) = 1 + x_{t-j}$ we will get $\lambda_j = \delta_j(1 + x_{t-j})$, this is the second example in their paper.

In the following we will show how to use Dirichlet process to study the BMTD model with re-parameterizations as shown in (8.3).

8.1 Gamma-Pareto Distribution

Here we will show a specific example of the general model above. This model is called Gamma-Pareto distribution. In model (8.2), if let $\alpha = \phi = 1$, $\delta = 0$, we will get k mixture of the Gamma-Pareto distribution with lag information

$$f_{x,y}(x, y) = \sum_{j=1}^{k} \pi_j \cdot \frac{x^{\gamma} \lambda_j^{\gamma} e^{-x\left(\lambda_j + \frac{y}{\beta_j}\right)}}{\Gamma(\gamma)\beta_j}.$$

As discussed before, let $\lambda_j = \frac{\theta_j}{x_{t-j}e^{-y_{t-j}}}$ and $\beta_j = \frac{1}{\beta_j}$, we have the re-parameterizations that Hassan and Lii (2006) used in their paper. Thus the BMTD model can be expressed as finite k mixtures as following

$$f(x_t, y_t | x^{t-1}, y^{t-1}) = \sum_{j=1}^{k} \pi_j \cdot \frac{x_t^{\gamma} \cdot \exp\left(-x_t\left(\frac{\theta_j}{x_{t-j}e^{-y_{t-j}}} + \beta_j y_t\right)\right)}{\left(\frac{1}{\theta_j} x_{t-j} e^{-y_{t-j}}\right)^{\gamma} \Gamma(\gamma) \frac{1}{\beta_j}}. \tag{8.4}$$

More generally, the model can be extended in the integral form as

$$\begin{aligned} & f(x_t, y_t | x^{t-1}, y^{t-1}) \\ & \quad = \int_{\mathcal{U}} \frac{x_t^{\gamma} \cdot \exp\left(-x_t\left(\frac{\theta_p}{x_{t-p}e^{-y_{t-p}}} + \beta_p y_t\right)\right)}{\left(\frac{1}{\theta_p} x_{t-p} e^{-y_{t-p}}\right)^{\gamma} \Gamma(\gamma) \frac{1}{\beta_p}} G(dp, d\theta_p, d\beta_p) \\ & \quad \triangleq \int_{\mathcal{U}} k(z_t | z^{t-1}, \mu_t) G(d\mu_t). \end{aligned} \tag{8.5}$$

Here $z_t = (x_t, y_t)$, $z^{t-1} = (x^{t-1}, y^{t-1}) = (x_{t-1}, \cdots, x_1; y_{t-1}, \cdots, y_1)$; $k(z_t|z^{t-1}, \mu_t)$ is the density function and $\mu_t = (p, \theta_p, \beta_p)$ is the parameters defined on the parameter space

$$\mathcal{U} = \{(p, \theta_p, \beta_p) : p \in \{1, 2, \cdots, K\}, K \geq 1 \text{ is a fixed integer}, \theta_p, \beta_p \in R\}.$$

K is the fixed constant which is the maximum of possible lag order.

This model (8.5) is an extension of the original model (8.4) by Hassan and Lii (2006). First, in model (8.4), the lag order in j^{th} component is j, which is fixed. However, in model (8.5), there is no such restriction, that means the lag order j now is a random variable and has no relation with the component order. Second, model (8.5) is an infinite mixtures rather than finite mixtures of model (8.4); This extension frees us from the need to pre-determine the number of mixtures k in (8.4).

8.2 Bayesian Analysis and Chinese Restaurant Process

In this example, the prior distribution of G is designed to be a Ferguson (1973) Dirichlet process $D(dG|\alpha, H)$. Here α is a positive constant, H is a probability measure defined on $(\mathcal{U}, \sigma(\mathcal{U}))$. From the definition of Dirichlet process, for each measurable partition $(\Theta_1, \cdots, \Theta_M)$ of $\mathcal{U}$, the vector $(G(\Theta_1), \cdots, G(\Theta_M))$ is distributed as a Dirichlet distribution with parameters $(\alpha H(\Theta_1), \cdots, \alpha H(\Theta_M))$. Under the prior D, the hierarchical form of a Bayesian infinite mixture of Gamma-Pareto distribution can be written as

$$\begin{aligned}
&z_t|z^{t-1}, \mu_t \sim k(z_t|z^{t-1}, \mu_t), \quad \text{for } t = 1, \cdots, n, \\
&\mu_t|G \overset{iid}{\sim} G(d\mu_t), \\
&G \sim D(dG|\alpha, H).
\end{aligned}$$

with $k(z_t|z^{t-1}, \mu_t)$ is the density function in (8.5).

8.2.1 Priors

To facilitate the calculation of posterior distribution, we will choose H with the following form: H has the called Discrete-Gamma-Gamma $(\rho_0, a_0, b_0, \lambda_0, r_0)$ distribution if $\rho_0 = (\rho_{0,1}, \cdots, \rho_{0,K})$ are the probabilities for the lag order p to take the possible values of 1 to K, i.e., $P(p = k) = \rho_{0,k}$. The priors for θ_p $(p = 1, \cdots, K)$ are the Gamma distributions with shape parameters $(a_{0,1}, \cdots, a_{0,K})$ and scale parameters $(b_{0,1}, \cdots, b_{0,K})$ respectively for all possible K orders. Given p, the prior distributions of β_p $(p = 1, \cdots, K)$ are Gamma distributions with scale parameters $(\lambda_{0,1}, \cdots, \lambda_{0,K})$

and shape parameters $(r_{0,1}, \cdots, r_{0,K})$ respectively. In summary, the priors will be set in the following forms

$$\begin{aligned}
&p \sim \text{Discrete}(\rho_{0,1}, \cdots, \rho_{0,K}) \quad p = 1, \cdots, K, \\
&\theta_p | p \sim \Gamma(a_{0,p}, b_{0,p}), \\
&\beta_p | p \sim \Gamma(r_{0,p}, \lambda_{0,p}).
\end{aligned}$$

So the joint distribution H is called Discrete-Gamma-Gamma distribution.

8.2.2 Posteriors

The Gibbs sampling scheme of Chinese Restaurant is used to generate partitions and samples from the posterior distribution.

Let $\mathbb{P} = \{T_1, \cdots, T_{n(p)}\}$ be an arbitrary partition of $\{1, 2, \cdots, n\}$ with size $n(p)$. Let e_j be the cardinality of each T_j for $j = 1, 2, \cdots, n(p)$. Denote the parameters μ in T_j by $(p_j, \theta_{p_j}, \beta_{p_j})$. In other words, for all the data in a subgroup T_j, the lag order is p_j, with the parameters θ_{p_j} and β_{p_j}.

Given table T_j, we can calculate the posterior distribution

$$\begin{aligned}
&p_j | T_j \sim \text{Discrete}(\rho_{j,1}, \rho_{j,2}, \cdots, \rho_{j,K}) \\
&\theta_{p_j} | p_j, T_j \sim \Gamma\left(e_j \gamma + a_{0,p_j}, \sum_{t \in T_j} \frac{x_t}{x_{t-p_j} e^{-y_{t-p_j}}} + b_{0,p_j} \right) \\
&\beta_{p_j} | p_j, T_j \sim \Gamma\left(e_j + r_{0,p_j}, \sum_{t \in T_j} x_t y_t + \lambda_{0,p_j} \right)
\end{aligned} \tag{8.6}$$

here

$$\begin{aligned}\rho_{j,k} =& \frac{\xi_{j,k}}{\sum_{k=1}^{K} \xi_{j,k}} \\ \xi_{j,k} =& \rho_{0,k} \cdot \prod_{t \in T_j} \left(\frac{x_t}{x_{t-k} e^{-y_{t-k}}} \right)^{\gamma} \cdot \left(\frac{1}{\Gamma(\gamma)} \right)^{e_j} \frac{\Gamma(e_j \gamma + a_{0,k})}{\Gamma(a_{0,k})} \frac{\Gamma(e_j + r_{0,k})}{\Gamma(r_{0,k})} \\ & \cdot \frac{(b_{0,k})^{a_{0,k}}}{\left(\sum\limits_{t \in T_j} \frac{x_t}{x_{t-k} e^{-y_{t-k}}} + b_{0,k} \right)^{e_j \gamma + a_{0,k}}} \cdot \frac{(\lambda_{0,k})^{r_{0,k}}}{\left(\sum\limits_{t \in T_j} x_t y_t + \lambda_{0,k} \right)^{e_j + r_{0,k}}}.\end{aligned}$$

See the appendix 8.7.2 for the proof of Equation (8.6).

8.3 Chinese Restaurant Partition Rules

Let $z = (x, y)$. Given the initial partition of $\{1, 2, \cdots, n\}$ as $\mathbb{P}^{(0)}$, one performs a cycle step from $\mathbb{P}^{(i)}$ to the next step $\mathbb{P}^{(i+1)}$ as follows: for each $t = 1, \cdots, n$, remove t from $\mathbb{P}^{(i)}$ to get a skip-t partition of $\{1, 2, \cdots, n\} - \{t\}$ denoted as $\mathbb{P}^{(i)}_{-t}$. Since customer t is removed, the new tables on skip-t partition $\mathbb{P}^{(i)}_{-t}$ is denoted as $\{T_{1,-t}, T_{2,-t}, \cdots\}$ Then reseat the customer t ($z_t = (x_t, y_t)$) on a new table or on an occupied table $T_{j,-t}$ of $\mathbb{P}^{(i)}_{-t}$ according to a specific probability: customer t is assigned to a new table with a probability proportional to

$$\frac{\alpha}{\alpha + n - 1} \int_{\mathcal{U}} k\left(z_t | z^{t-1}, \mu\right) G_0(d\mu) = \frac{\alpha}{\alpha + n - 1} \sum_{k=1}^{K} \zeta_{0,k}, \tag{8.7}$$

or to the table $T_{j,-t}$ with probability proportional to

$$\frac{e_{j,-t}}{\alpha + n - 1} \int_{\mathcal{U}} k\left(z_t | z^{t-1}, \mu\right) \pi(d\mu | T_{j,-t}) = \frac{e_{j,-t}}{\alpha + n - 1} \frac{\sum\limits_{k=1}^{K} \omega_{j,k}}{\sum\limits_{k=1}^{K} \xi_{j,k}} \tag{8.8}$$

where

$e_{j,-t}$ is the number of data points (observations) on table $T_{j,-t}$,

$$\begin{aligned}\zeta_{0,k} =& \rho_{0,k} \left(\frac{x_t}{x_{t-k} e^{-y_{t-k}}} \right)^{\gamma} \frac{1}{\Gamma(\gamma)} \frac{(b_{0,k})^{a_{0,k}}}{\Gamma(a_{0,k})} \frac{(\lambda_{0,k})^{r_{0,k}}}{\Gamma(r_{0,k})} \\ & \cdot \frac{\Gamma(\gamma + a_{0,k})}{\left(\frac{x_t}{x_{t-k} e^{-y_{t-k}}} + b_{0,k} \right)^{\gamma + a_{0,k}}} \cdot \frac{\Gamma(1 + r_{0,k})}{(x_t y_t + \lambda_{0,k})^{1 + r_{0,k}}}, \\ \omega_{j,k} =& \rho_{0,k} \cdot \left(\frac{x_0}{x_{0-k} e^{-y_{0-k}}} \right)^{\gamma} \left(\frac{1}{\Gamma(\gamma)} \right)^{e_j + 1} \\ & \cdot \prod_{t \in T_j} \left(\frac{x_t}{x_{t-k} e^{-y_{t-k}}} \right)^{\gamma} \cdot \frac{(b_{0,k})^{a_{0,k}}}{\Gamma(a_{0,k})} \frac{(\lambda_{0,k})^{r_{0,k}}}{\Gamma(r_{0,k})} \\ & \cdot \frac{\Gamma(e_j + r_{0,k} + 1)}{\left(x_0 y_0 + \sum_{t \in T_j} x_t y_t + \lambda_{0,k} \right)^{e_j + r_{0,k} + 1}} \\ & \cdot \frac{\Gamma(e_j \gamma + a_{0,k} + \gamma)}{\left(\frac{x_0}{x_{0-k} e^{-y_{0-k}}} + \sum_{t \in T_j} \frac{x_t}{x_{t-k} e^{-y_{t-k}}} + b_{0,k} \right)^{e_j \gamma + a_{0,k} + \gamma}}, \\ \xi_{j,k} =& \rho_{0,k} \cdot \prod_{t \in T_j} \left(\frac{x_t}{x_{t-k} e^{-y_{t-k}}} \right)^{\gamma} \cdot \left(\frac{1}{\Gamma(\gamma)} \right)^{e_j} \frac{\Gamma(e_j \gamma + a_{0,k})}{\Gamma(a_{0,k})} \frac{\Gamma(e_j + r_{0,k})}{\Gamma(r_{0,k})} \\ & \cdot \frac{(b_{0,k})^{a_{0,k}}}{\left(\sum_{t \in T_j} \frac{x_t}{x_{t-k} e^{-y_{t-k}}} + b_{0,k} \right)^{e_j \gamma + a_{0,k}}} \cdot \frac{(\lambda_{0,k})^{r_{0,k}}}{\left(\sum_{t \in T_j} x_t y_t + \lambda_{0,k} \right)^{e_j + r_{0,k}}}. \end{aligned}$$

The calculation of (8.7) and (8.8) is on appendix 8.7.3 and 8.7.4.

Based on this partition rules, we re-seat data t $z_t = (x_t, y_t)$ $(t = 1 \cdots n)$ (or we call it customer t) in the following way: first remove z_t from its current table, then it can sit on one of existing table with probability proportional to (8.8) or it may sit on a new table with probability proportional to (8.7). After we re-seat from $t = 1$ to $t = n$, we get a new partition $\mathbb{P}$. Then we calculate the posterior from 8.2.2 and sample the parameters from their posterior.

8.4 Algorithm

Now we can give the algorithm of Chinese Restaurant for the BMTD model (8.4). This algorithm tells us how to sample the parameters from their posterior distribution. As introduced before, after we get the samples, we can estimate the parameters from these samples.

Given the initial partition of $\{1, 2, \cdots, n\}$ denoted as $\mathbb{P}^{(0)}$, the algorithm gives iterations from partition $\mathbb{P}^{(i)}$ to partition $\mathbb{P}^{(i+1)}$.

- Sample $\mu^{(i)}$ (here μ is a vector, not a scalar) from the posterior distribution (8.6);
- For each $t = 1, 2, \cdots, n$, using (8.8) and (8.7) to determine the new table that customer t should sit;
- Implemented this for all n data points, then we will get the new partition $\mathbb{P}^{(i+1)}$;

Repeat these steps $M + N$ times. The first M data is treated as burning-in. The rest N sample data is used to estimate parameters.

From these N data points, denote the size of $\mathbb{P}^{(i)}$ as k_i and the corresponding parameters as $\mu^{(i)}$, $i = 1, \cdots, N$. Number of mixture components is estimated by the mode of $k_i's$. Suppose it is k_0, then we select all those $\mu^{(i)}$ that has k_0 components. The parameter vector μ is estimated by the average of these $\mu^{(i)}{}'s$.

8.5 Simulation Results

Here two simulations are given for the example (8.5). For these simulations, we will study how well does Dirichlet process perform in the BMTD model. To minimize the influence of the priors, we will let the prior parameters $a_0 = b_0 = r_0 = \lambda_0$ all equal to 1. We will choose different value of α to study the influence of α. Also, the impact of the sample size n is studied.

8.5.1 First Example

The true model is given by

$$\begin{aligned}
&f(x_t, y_t|x^{t-1}, y^{t-1}) = \\
&\quad .45 \times \frac{x_t^{2.5} \cdot \exp\left(-x_t\left(\frac{4.7}{x_{t-1}e^{-y_{t-1}}} + 3.8y_t\right)\right)}{\left(\frac{1}{4.7}x_{t-1}e^{-y_{t-1}}\right)^{2.5}\Gamma(2.5)\frac{1}{3.8}} \\
&\quad + .30 \times \frac{x_t^{2.5} \cdot \exp\left(-x_t\left(\frac{4.1}{x_{t-2}e^{-y_{t-2}}} + 3.3y_t\right)\right)}{\left(\frac{1}{4.1}x_{t-2}e^{-y_{t-2}}\right)^{2.5}\Gamma(2.5)\frac{1}{3.3}} \\
&\quad + .25 \times \frac{x_t^{2.5} \cdot \exp\left(-x_t\left(\frac{3.3}{x_{t-3}e^{-y_{t-3}}} + 2.8y_t\right)\right)}{\left(\frac{1}{3.3}x_{t-3}e^{-y_{t-3}}\right)^{2.5}\Gamma(2.5)\frac{1}{2.8}}
\end{aligned}$$

For this example, we study the BMTD model in which the lag order is determined by its mixture order; i.e., for the first component, the lag order is 1; for the second component, the lag order is 2; and so on. There are three components for the true model, that is, true $k = 3$. True value for the proportion of each component is $\pi_1 = .45$, $\pi_2 = .3$, $\pi_3 = .25$ and the corresponding lag orders are 1, 2 and 3 specifically. True value of $\theta_1 = 4.7$, $\theta_2 = 4.1$, $\theta_3 = 3.3$. True value of $\beta_1 = 3.8$, $\beta_2 = 3.3$, $\beta_3 = 2.8$. For each component, the maximum possible lag order K, is set to

10. We choose $a_0 = b_0 = r_0 = \lambda_0 = (1,1,1,1,1,1,1,1,1,1)$. The sample size n is set to $n = 100,\ 250$ and 400 respectively. Repeat the simulation 100 times. For each of these 100 replications, the MCMC is executed 20,000 times. The first 10,000 times are burning-in and the rest 10,000 data are used to estimate the parameters.

Dirichlet process $DP(\alpha, H)$ has α as a parameter. For α set to 2, 5 and 10, it is shown the influence is not significant from the simulation. In fact, from the simulation below we can almost ignore the influence of α. For the sample size n, we find the simulation result will be better as the sample size increases. When sample size is increased from 100 to 250 to 400, for 100 replications, the correct times of estimated number of mixture components k equaling to the true k increased from 92 to 100, also the correct times of lag order increase from 90 to 96 to 100. More details are given below.

Influence of sample size n and alpha

n	$\alpha = 2$		$\alpha = 5$		$\alpha = 10$	
	correct k	correct p	correct k	correct p	correct k	correct p
100	92	90	93	90	93	90
250	100	96	100	96	100	96
400	100	100	100	100	100	100

Here n is the number of mixture components, and p is the lag order

True Values			**Estimation of Parameters**			**STD error of Estimation**		
π	θ	β	π	θ	β	π	θ	β
0.45	4.7	3.8	0.4468	4.7021	3.8363	0.0328	0.0087	0.0201
0.30	4.1	3.3	0.2976	4.0603	3.3103	0.0103	0.0106	0.0709
0.25	3.3	2.8	0.2556	3.1868	2.8632	0.0606	0.0019	0.1032

The above simulation result based on sample size $n = 250$ and $\alpha = 2$. From above for $n = 250$, of the 100 replications, there are 100 times getting the correct number of k. Of these 100 correct k, furthermore there are 96 times we get correct lag order p. The estimation of parameters θ and β is estimated based on the 96 correct estimations.

8.5.2 Second Example

The true model is given by

$$
\begin{aligned}
&f(x_t, y_t|x^{t-1}, y^{t-1}) = \\
&\quad .32 \times \frac{x_t^{2.5} \cdot \exp\left(-x_t\left(\frac{4.5}{x_{t-1}e^{-y_{t-1}}} + 4.0y_t\right)\right)}{\left(\frac{1}{4.5}x_{t-1}e^{-y_{t-1}}\right)^{2.5}\Gamma(2.5)\frac{1}{4.0}} \\
&\quad + .23 \times \frac{x_t^{2.5} \cdot \exp\left(-x_t\left(\frac{3.8}{x_{t-2}e^{-y_{t-2}}} + 3.5y_t\right)\right)}{\left(\frac{1}{3.8}x_{t-2}e^{-y_{t-2}}\right)^{2.5}\Gamma(2.5)\frac{1}{3.5}} \\
&\quad + .30 \times \frac{x_t^{2.5} \cdot \exp\left(-x_t\left(\frac{3.3}{x_{t-3}e^{-y_{t-3}}} + 3.1y_t\right)\right)}{\left(\frac{1}{3.3}x_{t-3}e^{-y_{t-3}}\right)^{2.5}\Gamma(2.5)\frac{1}{3.1}} \\
&\quad + .15 \times \frac{x_t^{2.5} \cdot \exp\left(-x_t\left(\frac{2.7}{x_{t-1}e^{-y_{t-1}}} + 2.6y_t\right)\right)}{\left(\frac{1}{2.7}x_{t-1}e^{-y_{t-1}}\right)^{2.5}\Gamma(2.5)\frac{1}{2.6}}
\end{aligned}
$$

For this example, we study the extended BMTD model of which the lag order has no relation with the mixture order; i.e., for the first component, the lag order may be 1; for the second component, the lag order is 3; and for the third component, the lag order may be 1 again; and so on. There are four components for the true model, that is, true $k = 4$. True value for the proportion of each component is $\pi_1 = .32$, $\pi_2 = .23$, $\pi_3 = .30$ and $\pi_4 = .15$ and the corresponding lag orders are 1, 2, 3 and 1. True value of $\theta_1 = 4.5$, $\theta_2 = 3.8$, $\theta_3 = 3.3$ and $\theta_4 = 2.7$. True value of $\beta_1 = 4.0$, $\beta_2 = 3.5$, $\beta_3 = 3.1$ and $\beta_4 = 2.6$. For each component, the maximum possible lag order K, is set to 10. The priors parameters are still chosen as $a_0 = b_0 = r_0 = \lambda_0 = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$. The sample size n is set to equal to 100, 250 and 400 respectively. We repeat each simulation 100 times. For each of these 100 replications, the MCMC is executed 20,000 times. The first 10,000 times are burning-in and the rest 10,000 will be used to estimate the parameters.

Also, for different α equal to 2, 5 and 10, it is shown from the simulation their influence is not significant. For the sample size n, we see the simulation result will be better as the sample size increases. When sample size increases from 100 to 250 to 400, for 100 replications, the correct times of estimated number of mixture components k equaling to the true k increased from 86, 90 to 97 respectively, also the correct times of lag order increase from 72, 88 to 96. More details are given as below.

Influence of sample size and alpha

n	$\alpha = 2$		$\alpha = 5$		$\alpha = 10$	
	correct k	correct p	correct k	correct p	correct k	correct p
100	86	72	86	71	86	72
250	90	88	90	88	90	88
400	97	96	98	96	98	96

True Values			**Estimation of Parameters**			**STD error of Estimation**		
π	θ	β	π	θ	β	π	θ	β
0.32	4.5	4.0	0.3125	4.5103	4.0169	0.0164	0.0305	0.0314
0.23	3.8	3.5	0.2329	3.7639	3.5062	0.0306	0.0384	0.0426
0.30	3.3	3.1	0.3029	3.3104	3.1230	0.0439	0.1561	0.0673
0.15	2.7	2.6	0.1517	2.6763	2.5853	0.1028	0.0765	0.1001

The above simulation result based on sample size $n = 250$ and $\alpha = 2$. From above for $n = 250$, of the 100 replications, there are 90 times getting the correct number of k. Of these 90 times simulations with correct k, furthermore there are 88 times we get correct lag order p. The estimation of parameters θ and β is estimated based on the 88 correct estimations. From the table above we can see the estimations of the parameters are very close to the true parameters. This helps to prove that DP mixtures for BMTD model are very good.

8.6 Summary

From above it is shown that Dirichlet process is an effective way to study Bivariate Mixture Transition Distribution models. It works effectively in finding the true mixture numbers, the lag order and parameter estimation. Compared with EM algorithm which requires pre-determined mixture number k, there is no such requirements here. Also, unlike EM algorithm which depends on the initial value to find the maximum point, there is no initial value selection problem in Dirichlet process method. If there are several local maximum points, EM algorithm may converge to one of these local maximum point. In Dirichlet process method, we don't care about this because we sample from the posterior.

Another advantage of Dirichlet process method is that the mixture model expressed in this way is infinite mixtures rather than finite mixtures. Also, in this way, the lag order is not restricted to the mixture component order. These two make Dirichlet process more general than the EM algorithm method. Although Dirichlet process needs more time than EM algorithm for calculation, it requires less restriction for the models and it can deal with wider conditions than model (8.4) and the corresponding EM algorithm.

For Dirichlet process we need to pre-determine the value of α. Although this is an arbitrary job, it's usually set to a small value, such as 2, 5 or 10 in our example. It is shown from the simulation that the selected value of alpha doesn't affect too much. In fact, it is demonstrated that the value of alpha has little effect of our studies. Another way of study the effect of alpha is set alpha as a random variable and set up a distribution for it. This is studied by Lau and So (2008). In their studies it's also shown the value of alpha is not significant.

8.7 Proof of Equations

8.7.1 Prove of Equation (8.3)

Let's look at the j^{th} component density function

$$f_j(x_t, y_t) = C_j x_t^{\delta_j+\gamma_j+1/\phi_j-1} y_t^{\delta_j} \cdot e^{-x_t^{\alpha_j}\left(\lambda_j + y_t^{\phi_j}/\beta_j^{\phi_j}\right)},$$

$$C_j = \frac{\alpha_j \phi_j \lambda_j^{\delta_j/\alpha_j+\gamma_j/\alpha_j-\delta_j/\phi_j+1/(\alpha_j\phi_j)-1/\phi_j}}{\beta_j^{\delta_j+1}\Gamma(\frac{\delta_j+1}{\phi_j})\Gamma(\frac{\delta_j}{\alpha_j}+\frac{\gamma_j}{\alpha_j}-\frac{\delta_j}{\phi_j}+\frac{1}{\alpha_j\phi_j}-\frac{1}{\phi_j})}. \tag{8.9}$$

We will rewrite the density function as

$$\begin{aligned} f_j(x_t, y_t) &= C_j x_t^{\delta_j+\gamma_j+1/\phi_j-1} y_t^{\delta_j} \cdot e^{-x_t^{\alpha_j}(\lambda_j + y_t^{\phi_j}/\beta_j^{\phi_j})} \\ &= C_j x_t^{\delta_j+\gamma_j+1/\phi_j-1} y_t^{\delta_j} \cdot e^{-x_t^{\alpha_j}\lambda_j} \cdot e^{x_t^{\alpha_j} y_t^{\phi_j}/\beta_j^{\phi_j}}. \end{aligned}$$

If we use the re-parameterizations of (8.3)

$$\lambda_j = \theta_{j,1} \cdot g_1\left(x^{t-1}, y^{t-1}\right)$$

$$\beta_j = \theta_{j,2} \cdot g_2\left(x^{t-1}, y^{t-1}\right)$$

we will get

$$
\begin{aligned}
f_j(x_t, y_t) &= C_j x_t^{\delta_j+\gamma_j+1/\phi_j-1} y_t^{\delta_j} \cdot e^{-x_t^{\alpha_j}\lambda_j} \cdot e^{x_t^{\alpha_j} y_t^{\phi_j}/\beta_j^{\phi_j}} \\
&\propto \lambda_j^{\delta_j/\alpha_j+\gamma_j/\alpha_j-\delta_j/\phi_j+1/(\alpha_j\phi_j)-1/\phi_j} \beta_j^{-(\delta_j+1)} \cdot e^{-x_t^{\alpha}\lambda_j} \cdot e^{x_t^{\alpha} y_t^{\phi}/\beta_j^{\phi}} \\
&\propto \theta_{j,1}^{\delta_j/\alpha_j+\gamma_j/\alpha_j-\delta_j/\phi_j+1/(\alpha_j\phi_j)-1/\phi_j} \theta_{j,2}^{-(\delta_j+1)} \\
&\quad \cdot \exp\left\{-x_t^{\alpha_j} \cdot \theta_{j,1} \cdot g_1\left(x^{t-1}, y^{t-1}\right)\right\} \\
&\quad \cdot \exp\left\{x_t^{\alpha_j} y_t^{\phi_j} \left(\theta_{j,2} \cdot g_2\left(x^{t-1}, y^{t-1}\right)\right)^{-\phi_j}\right\} \\
&\propto \theta_{j,1}^{\delta_j/\alpha_j+\gamma_j/\alpha_j-\delta_j/\phi_j+1/(\alpha_j\phi_j)-1/\phi_j} \theta_{j,2}^{-(\delta_j+1)} \\
&\quad \cdot \exp\left\{-\theta_{j,1} \cdot x_t^{\alpha_j} \cdot g_1\left(x^{t-1}, y^{t-1}\right)\right\} \\
&\quad \cdot \exp\left\{\theta_{j,2} \cdot x_t^{\alpha_j} y_t^{\phi_j} \left(\theta_{j,2} \cdot g_2\left(x^{t-1}, y^{t-1}\right)\right)^{-\phi_j}\right\}.
\end{aligned}
$$

By looking at this equation, we see that the parameters $\theta_{j,1}$ and $\theta_{j,2}$ having the exponentially-like construction. So, if we choose the Gamma Distribution priors for $\theta_{j,1}$ and $\theta_{j,2}$, the posterior will still be exponential distribution. This explains why re-parametrization in (8.3) helps to find conjugate priors.

8.7.2 Proof of the Posterior Distribution (Equation (8.6))

Denote $\mu_j^* = (p_j, \theta_{p_j}, \beta_{p_j})$. Given the table T_j, the posterior distribution is proportional to the product of the likelihood in T_j and the prior H

$$
\begin{aligned}
&posterior\ of\ (p_j, \theta_{p_j}, \beta_{p_j}|T_j) \\
&\propto \prod_{t \in T_j} k\left((x_t, y_t)|x^{t-1}, y^{t-1}, \mu_j^*\right) \cdot H(d\mu_j^*) \\
&= \prod_{t \in T_j} \frac{x_t^\gamma \cdot \exp\left(-x_t\left(\frac{\theta_{p_j}}{x_{t-p_j}e^{-y_{t-p_j}}} + y_t\beta_{p_j}\right)\right)}{\left(\frac{1}{\theta_{p_j}}x_{t-p_j}e^{-y_{t-p_j}}\right)^\gamma \Gamma(\gamma)\frac{1}{\beta_{p_j}}} \\
&\quad \cdot \rho_{0,p_j} \cdot \frac{(b_{0,p_j})^{a_{0,p_j}}}{\Gamma(a_{0,p_j})}\theta_{p_j}^{a_{0,p_j}-1}e^{-b_{0,p_j}\theta_{p_j}} \cdot \frac{(\lambda_{0,p_j})^{r_{0,p_j}}}{\Gamma(r_{0,p_j})}\beta_{p_j}^{r_{0,p_j}-1}e^{-\lambda_{0,p_j}\beta_{p_j}} \\
&= \frac{\prod_{t \in T_j} x_t^\gamma \cdot \exp\left(-\theta_{p_j}\sum_{t \in T_j}\frac{x_t}{x_{t-p_j}e^{-y_{t-p_j}}} - \beta_{p_j}\sum_{t \in T_j} x_t y_t\right)}{\left(\frac{1}{\theta_{p_j}}\right)^{e_j\gamma} \cdot (\Gamma(\gamma))^{e_j} \cdot \left(\frac{1}{\beta_{p_j}}\right)^{e_j} \cdot \prod_{t \in T_j}\left(x_{t-p_j}e^{-y_{t-p_j}}\right)^\gamma} \\
&\quad \cdot \rho_{0,p_j} \cdot \frac{(b_{0,p_j})^{a_{0,p_j}}}{\Gamma(a_{0,p_j})}\theta_{p_j}^{a_{0,p_j}-1}e^{-b_{0,p_j}\theta_{p_j}} \cdot \frac{(\lambda_{0,p_j})^{r_{0,p_j}}}{\Gamma(r_{0,p_j})}\beta_{p_j}^{r_{0,p_j}-1}e^{-\lambda_{0,p_j}\beta_{p_j}} \\
&= \rho_{0,p_j} \cdot \prod_{t \in T_j}\left(\frac{x_t}{x_{t-p_j}e^{-y_{t-p_j}}}\right)^\gamma \cdot \left(\frac{1}{\Gamma(\gamma)}\right)^{e_j} \frac{(b_{0,p_j})^{a_{0,p_j}}}{\Gamma(a_{0,p_j})}\frac{(\lambda_{0,p_j})^{r_{0,p_j}}}{\Gamma(r_{0,p_j})} \\
&\quad \cdot (\theta_{p_j})^{e_j\gamma + a_{0,p_j}-1} \cdot \exp\left(-\theta_{p_j}\left(\sum_{t \in T_j}\frac{x_t}{x_{t-p_j}e^{-y_{t-p_j}}} + b_{0,p_j}\right)\right) \\
&\quad \cdot (\beta_{p_j})^{e_j + r_{0,p_j}-1} \cdot \exp\left(-\beta_{p_j}\left(\sum_{t \in T_j} x_t y_t + \lambda_{0,p_j}\right)\right)
\end{aligned}
$$

$$
\begin{aligned}
&=\rho_{0,p_j} \cdot \prod_{t\in T_j} \left(\frac{x_t}{x_{t-p_j} e^{-y_{t-p_j}}} \right)^{\gamma} \cdot \left(\frac{1}{\Gamma(\gamma)} \right)^{e_j} \frac{(b_{0,p_j})^{a_{0,p_j}}}{\Gamma(a_{0,p_j})} \frac{(\lambda_{0,p_j})^{r_{0,p_j}}}{\Gamma(r_{0,p_j})} \\
&\cdot \frac{\Gamma(e_j\gamma + a_{0,p_j})}{\left(\sum\limits_{t\in T_j} \frac{x_t}{x_{t-p_j} e^{-y_{t-p_j}}} + b_{0,p_j} \right)^{e_j\gamma + a_{0,p_j}}} \frac{\Gamma(e_j + r_{0,p_j})}{\left(\sum\limits_{t\in T_j} x_t y_t + \lambda_{0,p_j} \right)^{e_j + r_{0,p_j}}} \\
&\cdot f_1(\theta_{p_j}) \cdot f_1(\beta_{p_j}) \\
&=\rho_{0,p_j} \cdot \prod_{t\in T_j} \left(\frac{x_t}{x_{t-p_j} e^{-y_{t-p_j}}} \right)^{\gamma} \left(\frac{1}{\Gamma(\gamma)} \right)^{e_j} \frac{\Gamma(e_j\gamma + a_{0,p_j})}{\Gamma(a_{0,p_j})} \frac{\Gamma(e_j + r_{0,p_j})}{\Gamma(r_{0,p_j})} \\
&\cdot \frac{(b_{0,p_j})^{a_{0,p_j}}}{\left(\sum\limits_{t\in T_j} \frac{x_t}{x_{t-p_j} e^{-y_{t-p_j}}} + b_{0,p_j} \right)^{e_j\gamma + a_{0,p_j}}} \cdot \frac{(\lambda_{0,p_j})^{r_{0,p_j}}}{\left(\sum\limits_{t\in T_j} x_t y_t + \lambda_{0,p_j} \right)^{e_j + r_{0,p_j}}} \\
&\cdot f_1(\theta_{p_j}) \cdot f_1(\beta_{p_j}) \\
&\triangleq \xi_{j,p_j} \cdot f_1(\theta_{p_j}) \cdot f_1(\beta_{p_j}) \qquad (8.10)
\end{aligned}
$$

where

$$
\begin{aligned}
f_1(\theta_{p_j}) &= \Gamma \left(e_j\gamma + a_{0,p_j}, \sum_{t\in T_j} \frac{x_t}{x_{t-p_j} e^{-y_{t-p_j}}} + b_{0,p_j} \right) \\
f_1(\beta_{p_j}) &= \Gamma \left(e_j + r_{0,p_j}, \sum_{t\in T_j} x_t y_t + \lambda_{0,p_j} \right).
\end{aligned}
$$

We denote

$$
\psi_{j,p_j} = \rho_{0,p_j} \cdot \prod_{t\in T_j} \left(\frac{x_t}{x_{t-p_j} e^{-y_{t-p_j}}} \right)^{\gamma} \cdot \left(\frac{1}{\Gamma(\gamma)} \right)^{e_j} \frac{(b_{0,p_j})^{a_{0,p_j}}}{\Gamma(a_{0,p_j})} \frac{(\lambda_{0,p_j})^{r_{0,p_j}}}{\Gamma(r_{0,p_j})}.
$$

From (8.10) it is easy to find that the posterior for p is discrete, the posterior for θ_p and β_p are Gamma, as expressed in (8.6).

8.7.3 Probability of Sitting On An Existing Table (Equation (8.8))

To calculate the probability of sitting on an existing table (8.8), we need to calculate $\int_{\mathcal{U}} k\left(z_0|z^{t-1}, \mu_j^*\right) \pi(d\mu_j^*|T_j)$. Denote $\mu_j^* = (p_j, \theta_{p_j}, \beta_{p_j})$. From Lau and So (2008) it is easy to get that

$$\pi(d\mu_j^*|T_j) = \frac{\prod_{i\in T_j} k\left((x_i, y_i)|\vec{x}_{i-1}, \vec{y}_{i-1}, \mu_j^*\right) H(d\mu_j^*)}{\int_{\mathcal{U}} \prod_{i\in T_j} k\left((x_i, y_i)|x^{i-1}, \vec{y}_{i-1}, \mu_j^*\right) H(d\mu_j^*)}.$$

So the integral can be rewritten as

$$\begin{aligned}
&\int_{\mathcal{U}} k\left(z_0|z^{t-1}, \mu_j^*\right) \pi(d\mu_j^*|T_j) \\
&\quad = \frac{\int_{\mathcal{U}} k\left((x_0, y_0)|x^{t-1}, y^{t-1}, \mu_j^*\right) \prod_{i\in T_j} k\left((x_i, y_i)|x^{i-1}, y^{i-1}, \mu_j^*\right) H(d\mu_j^*)}{\int_{\mathcal{U}} \prod_{i\in T_j} k\left((x_i, y_i)|x^{i-1}, y^{i-1}, \mu_j^*\right) H(d\mu_j^*)} \\
&\quad \triangleq \frac{m(z_0 \cup T_j)}{m(T_j)}. \qquad (8.11)
\end{aligned}$$

We will calculate the nominator and denominator separately.

First we will calculate the denominator $m(T_j)$. Remember that from previous calculation we know

$$\prod_{i\in T_j} k\left((x_i, y_i)|x^{i-1}, y^{i-1}, \mu_j^*\right) H(d\mu_j^*) = \xi_{j,p_j} \cdot f_1(\theta_{p_j}) \cdot f_1(\beta_{p_j}).$$

So

$$m(T_j) = \int_{\mathcal{U}} \prod_{i\in T_j} k\left((x_i, y_i)|x^{i-1}, y^{i-1}, \mu_j^*\right) H(d\mu_j^*) = \sum_{k=1}^{K} \xi_{j,k}. \qquad (8.12)$$

Next to calculate the nominator part $m(z_0 \cup T_j)$.

$$m(z_0 \cup T_j) = \int_{\mathcal{U}} k\left((x_0, y_0)|x^{t-1}, y^{t-1}, \mu_j^*\right) \prod_{i\in T_j} k\left((x_i, y_i)|x^{i-1}, y^{i-1}, \mu_j^*\right) H(d\mu_j^*). \qquad (8.13)$$

$$
\begin{aligned}
& k\left((x_0,y_0)|x^{t-1},y^{t-1},\mu_j^*\right)\prod_{i\in T_j} k\left((x_i,y_i)|x^{i-1},y^{i-1},\mu_j^*\right)H(d\mu_j^*) \\
&= \frac{x_0^{\gamma}\cdot\exp\left(-x_0\left(\frac{\theta_{p_j}}{x_{0-p_j}e^{-y_{0-p_j}}}+\beta_{p_j}y_0\right)\right)}{\left(\frac{1}{\theta_{p_j}}x_{0-p_j}e^{-y_{0-p_j}}\right)^{\gamma}\Gamma(\gamma)\frac{1}{\beta_{p_j}}}\xi_{j,p_j}f_1(\theta_{p_j})f_1(\beta_{p_j}) \\
&= \left(\frac{x_0}{x_{0-p_j}e^{-y_{0-p_j}}}\right)^{\gamma}\frac{1}{\Gamma(\gamma)}\left(\theta_{p_j}\right)^{\gamma}\cdot\exp\left(-\theta_{p_j}\frac{x_0}{x_{0-p_j}e^{-y_{0-p_j}}}\right) \\
&\quad\cdot\beta_{p_j}e^{-\beta_{p_j}x_0y_0}\cdot\psi_{j,p_j} \\
&\quad\cdot\left(\theta_{p_j}\right)^{e_j\gamma+a_{0,p_j}-1}\cdot\exp\left(-\theta_{p_j}\left(\sum_{i\in T_j}\frac{x_i}{x_{i-p_j}e^{-y_{i-p_j}}}+b_{0,p_j}\right)\right) \\
&\quad\cdot\left(\beta_{p_j}\right)^{e_j+r_{0,p_j}-1}\cdot\exp\left(-\beta_{p_j}\left(\sum_{i\in i_j}x_iy_i+\lambda_{0,p_j}\right)\right) \\
&= \left(\frac{x_0}{x_{0-p_j}e^{-y_{0-p_j}}}\right)^{\gamma}\frac{1}{\Gamma(\gamma)}\psi_{j,p_j} \\
&\quad\cdot\left(\theta_{p_j}\right)^{e_j\gamma+a_{0,p_j}+\gamma-1}\cdot\exp\left(-\theta_{p_j}\left(\frac{x_0}{x_{0-p_j}e^{-y_{0-p_j}}}+\sum_{i\in T_j}\frac{x_i}{x_{i-p_j}e^{-y_{i-p_j}}}+b_{0,p_j}\right)\right) \\
&\quad\cdot\left(\beta_{p_j}\right)^{e_j+r_{0,p_j}+1-1}\cdot\exp\left(-\beta_{p_j}\left(x_0y_0+\sum_{i\in T_j}x_iy_i+\lambda_{0,p_j}\right)\right) \\
&= \left(\frac{x_0}{x_{0-p_j}e^{-y_{0-p_j}}}\right)^{\gamma}\frac{1}{\Gamma(\gamma)}\psi_{j,p_j} \\
&\quad\cdot\frac{\Gamma(e_j\gamma+a_{0,p_j}+\gamma)}{\left(\frac{x_0}{x_{0-p_j}e^{-y_{0-p_j}}}+\sum_{i\in T_j}\frac{x_i}{x_{i-p_j}e^{-y_{i-p_j}}}+b_{0,p_j}\right)^{e_j\gamma+a_{0,p_j}+\gamma}} \\
&\quad\cdot\frac{\Gamma(e_j+r_{0,p_j}+1)}{\left(x_0y_0+\sum_{i\in T_j}x_iy_i+\lambda_{0,p_j}\right)^{e_j+r_{0,p_j}+1}} \\
&\quad\cdot f_2(\theta_{p_j})f_2(\beta_{p_j}) \\
&\triangleq \omega_{j,p_j}\cdot f_2(\theta_{p_j})f_2(\beta_{p_j}) \qquad (8.14)
\end{aligned}
$$

where

$$
\begin{aligned}
&\omega_{j,p_j} \qquad (8.15)\\
&=\left(\frac{x_0}{x_{0-p_j}e^{-y_{0-p_j}}}\right)^{\gamma}\frac{1}{\Gamma(\gamma)}\psi_{j,p_j}\\
&\cdot\frac{\Gamma(e_j\gamma+a_{0,p_j}+\gamma)}{\left(\frac{x_0}{x_{0-p_j}e^{-y_{0-p_j}}}+\sum_{t\in T_j}\frac{x_t}{x_{t-p_j}e^{-y_{t-p_j}}}+b_{0,p_j}\right)^{e_j\gamma+a_{0,p_j}+\gamma}}\\
&\cdot\frac{\Gamma(e_j+r_{0,p_j}+1)}{\left(x_0y_0+\sum_{t\in T_j}x_ty_t+\lambda_{0,p_j}\right)^{e_j+r_{0,p_j}+1}}\\
&=\rho_{0,p_j}\cdot\left(\frac{x_0}{x_{0-p_j}e^{-y_{0-p_j}}}\right)^{\gamma}\left(\frac{1}{\Gamma(\gamma)}\right)^{e_j+1}\prod_{t\in T_j}\left(\frac{x_t}{x_{t-p_j}e^{-y_{t-p_j}}}\right)^{\gamma}\\
&\cdot\frac{(b_{0,p_j})^{a_{0,p_j}}}{\Gamma(a_{0,p_j})}\frac{(\lambda_{0,p_j})^{r_{0,p_j}}}{\Gamma(r_{0,p_j})}\\
&\cdot\frac{\Gamma(e_j+r_{0,p_j}+1)}{\left(x_0y_0+\sum_{t\in T_j}x_ty_t+\lambda_{0,p_j}\right)^{e_j+r_{0,p_j}+1}}\\
&\cdot\frac{\Gamma(e_j\gamma+a_{0,p_j}+\gamma)}{\left(\frac{x_0}{x_{0-p_j}e^{-y_{0-p_j}}}+\sum_{t\in T_j}\frac{x_t}{x_{t-p_j}e^{-y_{t-p_j}}}+b_{0,p_j}\right)^{e_j\gamma+a_{0,p_j}+\gamma}},
\end{aligned}
$$

$$
\begin{aligned}
f_2(\theta_{p_j})&=\Gamma\left(e_j\gamma+a_{0,p_j}+\gamma,\frac{x_0}{x_{0-p_j}e^{-y_{0-p_j}}}+\sum_{t\in T_j}\frac{x_t}{x_{t-p_j}e^{-y_{t-p_j}}}+b_{0,p_j}\right),\\
f_2(\beta_{p_j})&=\Gamma\left(e_j+r_{0,p_j}+1,x_0y_0+\sum_{t\in T_j}x_ty_t+\lambda_{0,p_j}\right).
\end{aligned}
$$

So,

$$
\begin{aligned}
&m(z_0 \cup T_j) \\
&\quad = \int_{\mathcal{U}} k\left((x_0, y_0)|x^{t-1}, y^{t-1}, \mu_j^*\right) \prod_{i \in T_j} k\left((x_i, y_i)|x^{i-1}, y^{i-1}, \mu_j^*\right) H(d\mu_j^*) \\
&\quad = \sum_{k=1}^{K} \omega_{j,k}. \qquad (8.16)
\end{aligned}
$$

Based on (8.12) and (8.16), we can get that the probability the data point $z_0 = (x_0, y_0)$ sitting on an existing table T_j equals to

$$
\frac{e_{j,-t}}{\alpha + n - 1} \int_{\mathcal{U}} k\left(z_0|z^{t-1}, \mu_j^*\right) \pi(d\mu_j^*|T_j) = \frac{e_{j,-t}}{\alpha + n - 1} \frac{\sum\limits_{k=1}^{K} \omega_{j,k}}{\sum\limits_{k=1}^{K} \xi_{j,k}}. \qquad (8.17)
$$

8.7.4 Probability of Sitting On A New Table (Equation (8.7))

The probability of data point $z_0 = (x_0, y_0)$ sitting on a new table is $\int_{\mathcal{U}} k\left(z_0|z^{t-1}, \mu\right) H(d\mu)$.

$$
\begin{aligned}
&k\left(z_0|z^{t-1}, \mu\right) H(d\mu) \\
&\quad = \frac{x_0^{\gamma} \cdot \exp\left(-x_0\left(\frac{\theta_p}{x_{0-p}e^{-y_{0-p}}} + \beta_p y_0\right)\right)}{\left(\frac{1}{\theta_p} x_{0-p} e^{-y_{0-p}}\right)^{\gamma} \Gamma(\gamma) \frac{1}{\beta_p}} \\
&\qquad \cdot \rho_{0,p} \cdot \frac{(b_{0,p})^{a_{0,p}}}{\Gamma(a_{0,p})} \theta_p^{a_{0,p}-1} e^{-b_{0,p}\theta_p} \cdot \frac{(\lambda_{0,p})^{r_{0,p}}}{\Gamma(r_{0,p})} \beta_p^{r_{0,p}-1} e^{-\lambda_{0,p}\beta_p} \\
&\quad = \left(\frac{x_0}{x_{0-p}e^{-y_{0-p}}}\right)^{\gamma} \frac{1}{\Gamma(\gamma)} \rho_{0,p} \frac{(b_{0,p})^{a_{0,p}}}{\Gamma(a_{0,p})} \frac{(\lambda_{0,p})^{r_{0,p}}}{\Gamma(r_{0,p})} \\
&\qquad \cdot \frac{\Gamma(\gamma + a_{0,p})}{\left(\frac{x_0}{x_{0-p}e^{-y_{0-p}}} + b_{0,p}\right)^{\gamma + a_{0,p}}} \cdot \frac{\Gamma(1 + r_{0,p})}{(x_0 y_0 + \lambda_{0,p})^{1 + r_{0,p}}} \cdot f_3(\theta_p) f_3(\beta_p)
\end{aligned}
$$

where

$$
\begin{aligned}
f_3(\theta_p) &= \Gamma\left(\frac{x_0}{x_{0-p}e^{-y_{0-p}}} + b_{0,p}, \gamma + a_{0,p}\right), \\
f_3(\beta_p) &= \Gamma(x_0 y_0 + \lambda_{0,p}, 1 + r_{0,p}).
\end{aligned}
$$

Based on these we get

$$
\begin{aligned}
&\int_{\mathcal{U}} k\left(z_t|z^{t-1}, \mu\right) H(d\mu) \\
&\quad = \sum_{i=1}^{K} \rho_{0,k} \left(\frac{x_t}{x_{t-k}e^{-y_{t-k}}}\right)^{\gamma} \frac{1}{\Gamma(\gamma)} \frac{(b_{0,k})^{a_{0,k}}}{\Gamma(a_{0,k})} \frac{(\lambda_{0,k})^{r_{0,k}}}{\Gamma(r_{0,k})} \\
&\qquad \cdot \frac{\Gamma(\gamma + a_{0,k})}{\left(\frac{x_t}{x_{t-k}e^{-y_{t-k}}} + b_{0,k}\right)^{\gamma + a_{0,k}}} \cdot \frac{\Gamma(1 + r_{0,k})}{(x_t y_t + \lambda_{0,k})^{1 + r_{0,k}}} \\
&\quad = \sum_{k=1}^{K} \zeta_{0,k}. \qquad (8.18)
\end{aligned}
$$

Chapter 9

Conclusion

9.1 Summary of My Research

In this thesis we focus on the questions from MTD/BMTD models. We have discussed the problem of singularity and the shortcomings of EM algorithm in Chapter 2. Then we discuss how to solve the problem of singularity under Bayesian framework.

Under Bayesian framework, we use two methods to estimate parameters after we get the posterior density. The first method is EM algorithm (Chapter 3 and Chapter 4). Although this EM algorithm method has its shortcomings, it gives us consistent estimates and we have proved the consistency in Chapter 5. The second method is to use MCMC to sample from the posterior of the parameters. After we get the posterior, we sample the values of parameters from their posterior and then estimate the parameters through these samples. For MCMC method, we have discussed two sampling methods. One is called Birth-Death process method (Chapter 6). In this method we treat the number of mixture components k as a random variable rather than a fixed number in EM algorithm. Another method is called Dirichlet process method (Chapter 7 and Chapter 8). This method is more flexible than EM algorithm and Birth-Death process. In Dirithlet process method, we can treat the lag order in MTD/BMTD models as a random variable. In each chapter we use simulations to demonstrate the advantages of Bayesian methods.

9.2 Future Works

There are some problems that we may continue to plan to investigate. In this thesis we use conjugate priors for the parameters in the models. However, we are not restricted to use conjugate priors. Neal (2000) gave several methods to sample from the posterior of normal mixture models when the conjugate priors were difficult to obtain. He introduced one Gibbs sampling method by using a set of auxiliary parameters. With his method, we can take non-conjugate priors for the Dirichlet process mixtures of MTD/BMTD models. This gives us more flexibility to choose priors.

Another interesting work that we can use is hierarchical Dirichlet process mixtures introduced by Y. W. Teh (2006). In Dirichlet process $DP(\alpha, H)$, H is called base measure on the parameter space Θ. In our study we also choose this H having a conjugate form (see 8.2.1) since this help us to get the posterior. Teh (2006) introduced the hierarchical Dirichlet process where H is distributed according to another Dirichlet process. If we use hierarchical Dirichlet process method, then H is not restricted to conjugate form. This method also gives us more flexibility to study MTD/BMTD models.

Bibliography

[1] Baudry, J.P., Raftery, A.E., Celeux, G., Lo, K. and Gottardo, R. (2010) Combining Mixture Components for Clustering. Journal of Computational and Graphical Statistics 19:332-353

[2] Berchtold, A. (2003) Mixture Transition Distribution (MTD) Modeling of Heteroscedastic Time Series. Computational Statistics and Data Analysis, 41 (3-4), 399-411

[3] Berchtold, A., Raftery, A. (2002) The Mixture Transition Distribution Model for High-Order Markov Chains and Non-Gaussian Time Series. Statistical Science Vol. 17, No. 3, 328C356

[4] Bishop, M. C. (2006) Mixture Models and the EM Algorithm. 2006 Advanced Tutorial Lecture Series, CUED

[5] Blackwell, D., and MacQueen, J. B. (1973) Ferguson Distributions via Polya urn Schemes, The Annals of Statistics, 1, 353-355

[6] Chanda, K. C. (1954) A note on the consistency and maxima ofthe roots ofthe likelihood equations. Biometrika 41, 56C61

[7] Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977) Maximum-likelihood from incomplete data via the em algorithm. J. Royal Statist. Soc. Ser. B., 39

[8] Farewell, V.T. (1982) The use of mixture models for the analysis of survival data with long-term survivors. Biometrics, 38, 1041-1046.

[9] Ferguson, T. S. (1973) A Bayesian analysis of some nonparametric problems. Annals of Statistics, 1:209-230

[10] Fraley, C. and Raftery, A.E (2007) Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering. Journal of Classification, 24, 155-181

[11] Gabriela Ciuperca, Andrea Ridolfi, Jerome Idier. (2000) Penalized Maximum Likelihood Estimator for Normal Mixtures

[12] Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82 711C732.

[13] Hathaway, R. J. (1985) A constrained formulation of maximum-likelihood estimation for normal mixture distributions. Ann. Statist. 13, 795C800

[14] Hassan, Y. M., Lii, Keh-shin. (2006) Modeling marked point process via Bivariate Mixture Transition Distribution Models, JASA, 1241-1252

[15] Khalid, El-Arini. (2008) Dirichlet ProcessesA gentle tutorial, Select Lab Meeting

[16] Lau, W. John, So, Mike. (2008) Bayesian mixture of autoregressive models, Computational Statistics and Data Analysis, 53, 38-60

[17] Le, N., Martin, R., and Raftery, E. (1996) Modeling Flat Stretches, Bursts, and-Outliers in Time Series Using Mixture Transition Distribution Models, JASA, 91, 1504-1515

[18] Lo, A.Y. (1984) On a class of bayesian nonparametric estimates: I. density estimates. Annals of Statistics, 12(1):351-357

[19] Lo, A.Y., Burner, L.J., Chan, A.T., (1996) Weighted Chinese Restaurant Processes and Bayesian mixture models, Research Report, HKUST

[20] MacEachern, S. N. and Muller, P. (1998) Estimating mixture of dirichlet process models. Journal of Computational and Graphical Statistics, 7:223C238

[21] McLachlan, G. J. and Peel, D. (2000) Finite mixture models. Wiley, New York.

[22] Neal, R. M. (1993) Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto

[23] Neal, R.M., (2000) Markov chain sampling methods for Dirichlet process process mixture models, Journal of Comptutational and Graphical Statistics, 9, 249-265

[24] Preston, C.J., (1976) Spatial birtd-and-death processes, Bulletin of the Institute of International Statistics, 46, 371-391

[25] Raftery, A.E. (1985) A model for high-order Markov chains. Journal of the Royal Statistical Society, series B, 47, 528-539

[26] Rasmussen, C.E. (2000) The infinite Gaussian mixture model. In Advances in Neural Information Processing Systems, volume 12

[27] Redner, R. (1980) Maximum likelihood estimation for mixture models. Technical memorandum, NASA.

[28] Redner, R. (1981) Note on the consistency ofthe maximum likelihood estimate for non-identifiable distributions. Ann. Statist. 9, 225C228

[29] Redner, R. A., Walker H. F. (1984) Mixture densities, maximum likelihood and the EM algorithm. SIAM Rev. 26, 195C239

[30] Richard, S., Green, J. (1997) On Bayesian Analysis of Mixture with an Unknown Number of Components, J. R. Statist. Soc. (B), 59, 731-792

[31] Ridolfi, A., Idier, J. (2000) Penalized maximum likelihood estimation for univariate normal mixture distributions. Bayesian inference and maximum entropy methods, MaxEnt Workshops. Gif-sur-Yvette, France, July 2000

[32] Ripley, B.D., (1977) Modeling Spatial Patterns (With Discussion), J. R. Statist. Soc. (B), 39, 172-212

[33] Rubin, Donald B., Gelman, Andrew, Stern, Hal (2003) Bayesian Data Analysis (2nd ed.). Boca Raton: Chapman and Hall/CRC.

[34] Sethuraman, J. (1994) A constructive defnition of Dirichlet priors. Statistica Sinica, 4:639-650

[35] Stephens, M. (1997) Bayesian methods for mixtures of normal distributions. PhD Thesis, University of Oxford.

[36] Stephens, M., (2000) Bayesian Analysis of Mixture Models with An Unknown Number of Components—An Alternative to Reversible Jump Method, The Annalas of Statistics, Vol. 28, 40-74

[37] Titterington, D.M., Smith, A.F., and Makov, U.E. (1985) Statistical Analysis of Finite Mixture Distributions, New York: Wiley.

[38] Wald, A. (1949) Note on the consistency ofthe maximum likelihood estimate. Ann. Math. Statist. 20, 595C601

[39] Wolfowitz, J. (1949) On Walds proof of the consistency of the maximum likelihood estimate. Ann. Math. Statist. 20, 601C602

[40] Wong,Chun Shan and Li, Wai Keung. (2000) On a mixture autoregressive model, J. R. Statist. Soc. B, 62, Part 1, 95-115

[41] Xu, L and Jordan, M. I., (1996) On Convergence Properties of the EM Algorithm for Gaussian Mixtures, Neural Computation 8: 129C151

[42] Y. W. Teh. (2006) A hierarchical Bayesian language model based on Pitman-Yorprocesses. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pages 985-992

[43] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. (2006) Hierarchical Dirichlet processes. Journal of the American Statistical Association, 101(476):1566-1581