

University of Kentucky

UKnowledge

Theses and Dissertations--Statistics

Statistics


2022

BETA MIXTURE AND CONTAMINATED MODEL WITH CONSTRAINTS AND APPLICATION WITH MICRO-ARRAY DATA

Ya Qi

University of Kentucky, qiya1989@gmail.com

Author ORCID Identifier:

 <https://orcid.org/0000-0001-6656-7008>

Digital Object Identifier: <https://doi.org/10.13023/etd.2022.315>

[Right click to open a feedback form in a new tab to let us know how this document benefits you.](#)

Recommended Citation

Qi, Ya, "BETA MIXTURE AND CONTAMINATED MODEL WITH CONSTRAINTS AND APPLICATION WITH MICRO-ARRAY DATA" (2022). *Theses and Dissertations--Statistics*. 64.

https://uknowledge.uky.edu/statistics_etds/64

This Doctoral Dissertation is brought to you for free and open access by the Statistics at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Statistics by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Ya Qi, Student

Dr. Richard Charnigo, Major Professor

Dr. Katherine Thompson, Director of Graduate Studies

BETA MIXTURE AND CONTAMINATED MODEL WITH CONSTRAINTS
AND APPLICATION WITH MICRO-ARRAY DATA

DISSERTATION

A dissertation submitted in partial
fulfillment of the requirements for
the degree of Doctor of Philosophy
in the College of Arts and Sciences
at the University of Kentucky

By

Ya Qi

Lexington, Kentucky

Director: Richard Charnigo, Professor of Statistics

Lexington, Kentucky

2022

Copyright© Ya Qi 2022

<https://orcid.org/0000-0001-6656-7008>

ABSTRACT OF DISSERTATION

BETA MIXTURE AND CONTAMINATED MODEL WITH CONSTRAINTS AND APPLICATION WITH MICRO-ARRAY DATA

This dissertation research is concentrated on the Contaminated Beta(CB) model and its application in micro-array data analysis. Modified Likelihood Ratio Test (MLRT) introduced by [Chen et al., 2001] is used for testing the omnibus null hypothesis of no contamination of Beta(1,1)([Dai and Charnigo, 2008]). We design constraints for two-component CB model, which put the mode toward the left end of the distribution to reflect the abundance of small p-values of micro-array data, to increase the test power. A three-component CB model might be useful when distinguishing high differentially expressed genes and moderate differentially expressed genes. If the null hypothesis above is rejected, we considered developing a method of testing the hypothesis of two-component vs three-component CB model. We first study CB model with one-parameter kernel distribution by fixing the other shape parameter across all the components. Using MLRT introduced by [Chen et al., 2004], we find the feasibility of this model after investigation. Then we consider a three-component CB model and designed constraints to guarantee the identifiability. We also study model selection and use sBIC introduced by [Drton and Plummer, 2017] to determine the number of components. We applied our tests and model to a toddler Down Syndrome data sets.

KEYWORDS: Micro-array data analysis, Finite mixture model, Contamination Beta mixture model, modified log-likelihood ratio test, model selection

Ya Qi

August 5, 2022

BETA MIXTURE AND CONTAMINATED MODEL WITH CONSTRAINTS
AND APPLICATION WITH MICRO-ARRAY DATA

By
Ya Qi

Richard Charnigo

Director of Dissertation

Katherine Thompson

Director of Graduate Studies

August 5, 2022

Date

ACKNOWLEDGMENTS

First of all, Please allow me to express my sincere appreciation to my advisors Dr. Richard Charnigo for all his help. My work and dissertation would be impossible if it is without his detailed guidance and persistent encouragement. He is the best advisor, I feel honored to be one of his students. I have been suffering from depression for two years. When I went through the tough time, Dr. Charnigo gives me a lot of support when I feel like giving up. He provide valuable suggestions and great encouragement during my dissertation research. He is an amazing advisor not just with profound knowledge and professional skill, but also has extraordinary patience and kindness.

Secondly, I would like to express my gratitude to Dr. Katherine Thompson. she give academic guidance during my study in the Department of Statistics and also give me a lot of support during my hard time. When I gave birth to my baby, she offer a nursing room the first time I back to school. When I suffered the depression and lost communication, she made effort to contact me and had a zoom meeting to encourage me. She also helped a lot during my dissertation exam application.

Thirdly, I would like to thank Dr. Arnold Stromberg. I learned a lot, especially communication and professional skills from him during the time I worked in the Applied Statistical Lab. It benefit me immeasurably.

Thirdly, I would like to thank Dr. Derek Young. He gave valuable technical suggestions, it is very helpful for my dissertation research and further study.

Fourthly, I would like to say thank you to my committee members: Dr. David

Fardo and Dr. David Jensen for their support and time.

Fifthly, I would like to say thank you to Zhen Zhang, my cousin work in Institute of Computing Technology , Chinese Academy of Science. She provide super computer for my real data analysis.

I also would like to thank the Department of Statistics for providing the financial support over the years and their strong backing to me all the time.

At the same time, I would like to express my special thanks to my Husband Dr. Li Xu. He encourage me to overcome any obstacles in finishing this dissertation. Thank my parents support me to fight the depression. Also my two kids Tiffany and Leon are my source of happiness, who can cheer me up whenever I need a lighter mood. I cannot do this without their love and supports for sure.

Last but not least, I would like to thank all the professors, friends, classmates that helped me through my Ph.D. studies. I will never forget this piece of precious journey for my whole life.

CONTENTS

Acknowledgments	iii
List of Tables	viii
List of Figures	ix
Chapter 1 Introduction	1
1.1 Finite Mixture Model	1
1.1.1 Definition of finite mixture model	1
1.1.2 Application of finite mixture model	2
1.2 Contaminated Density Model	3
1.3 Zero-one-inflated beta model	6
1.4 Estimate number of components by Testing or Information Criteria	7
1.5 EM Algorithm	12
1.5.1 E-step	13
1.5.2 M-step	13
Chapter 2 Two-component Beta Mixture Model with Constraints	15
2.1 Introduction	15
2.2 Estimate the MMLEs with constraints	16
2.3 The homogeneity hypothesis testing	18
2.3.1 The null limiting distribution	20
2.4 Simulation study	22
2.4.1 Actual rejection rate	22
2.4.2 Power	24
2.4.3 Interpolation with Small sample	30

2.5	Real data application	33
2.5.1	introduction to the data	33
2.5.2	Results	37
Chapter 3	Three-component Beta Mixture Model without constraints	42
3.1	Introduction	42
3.2	Hypothesis Test	44
3.3	Proof of conditions	45
3.4	Simulation study	48
3.4.1	Null distributions	49
3.4.2	Alternative distributions	49
3.4.3	Actual rejection rates and powers	51
3.4.4	Problems in MLRT with one parameter family	58
3.5	Real Data Application	61
3.5.1	Introduction	61
3.5.2	Results	64
Chapter 4	Three-component Beta Mixture Model with Constraints	69
4.1	Introduction	69
4.2	Identifiability of Beta Mixture Models	70
4.2.1	Identification of Three-component Beta Mixture Model	70
4.2.2	Identifiability of Three-component Contaminated Beta Model	75
4.2.3	Identifiability of Two-component Beta Mixture model	77
4.3	Estimating the MLEs	80
4.4	Hypothesis Testing	82
4.5	Introduction of sBIC	84
4.6	Simulation	89
4.6.1	Null distribution	89

4.6.2	Alternative distributions	89
4.7	Real Data Application	92
4.7.1	Introduction of real data	92
4.7.2	Results	92
Chapter 5	Summary and future Work	100
	Bibliography	103
	Vita	108
Ya Qi	108
EDUCATION	108
PROFESSIONAL EXPERIENCE	108
PUBLICATIONS AND PRESENTATIONS	108

LIST OF TABLES

2.1	The Characteristic of the simulated MLRT with constraints($n=2000$) . . .	20
2.2	$0.75\text{Beta}(1,1)+0.25\text{Beta}(0.7,2)$	24
2.3	$0.9\text{Beta}(1,1)+0.1\text{Beta}(0.5,1.5)$	26
2.4	$0.95\text{Beta}(1,1)+0.05\text{Beta}(0.5,1.5)$	27
2.5	$0.95\text{Beta}(1,1)+0.05\text{Beta}(0.6,3)$	28
2.6	Examples of fitted Beta contamination model	41
4.1	MLEs of some fitted constrained model in data 4	98

LIST OF FIGURES

1.1	An example of two-component normal mixture model	2
2.1	Basic shape of contamination Beta mixture models (modified from [Dai and Charnigo, 2008])	17
2.2	The Characteristics of MLRT vs simple size	19
2.3	Histogram of simulated MLRT with constraints ($n=2000$)	21
2.4	Actual rejection rates vs sample size	23
2.5	Power curve vs sample size	25
2.6	Power curves vs sample size of (V)-(VI)	29
2.7	Power curves vs sample size of (VII)-(X)	30
2.8	Compare MMLEs and true parameter	31
2.9	Shape of scenario (V) – (X)	32
2.10	Estimate critical value when sample size is small	33
2.11	Power curves when sample size is small	34
2.12	Histogram of all p-values, $n=461258$	35
2.13	Histogram of p-values on chromosome 21	36
2.14	Histogram of p-values of Data 3	37
2.15	Histogram of p-values of Data 4 (6 example)	40
3.1	Actual rejection rate when $\beta=1$	52
3.2	Actual rejection rate when $\alpha=1$	53
3.3	Power curves when $\beta=1$	54
3.4	power curves when $\alpha=1$	55
3.5	Basic Shape of two-component Beta contamination model	56
3.6	Power curves of sample generate from $0.5\text{Beta}(1,1)+0.5\text{Beta}(0.5,3)$	57
3.7	Density plot of (A12) and (A15)	57

3.8	The Histogram and fitted model when fix $\alpha=1$, the red line show 2-component CB model, the green line show 3-component CB model	58
3.9	The Histogram and fitted model when fix $\beta=1$, the red line show 2-component CB model, the green line show 3-component CB model	59
3.10	Histogram of p-values in Data 1 with $\alpha =1$, the red line shows the two-component fitted model, and the green line shows the three-component fitted model.	62
3.11	Histogram of p-values in Data 1 with $\beta =1$, the red line shows the two-component fitted model, and the green line shows the three-component fitted model.	63
3.12	Histogram of p-values in Data 2 with $\beta =1$, the red line shows the two-component fitted model, and the green line shows the three-component fitted model.	66
3.13	Histogram of p-values in Data 3 with $\beta =1$, the red line shows the two-component fitted model, and the green line shows the three-component fitted model.	68
4.1	Histogram of LRT when sample size $n=1000$	83
4.2	Percentile plots of LRT statistic under H_0 vs sample size	84
4.3	Percentile plots of LRT statistic under H_0 vs sample size	85
4.4	Percentile plots of LRT statistic under H_0 vs sample size	86
4.5	Actual rejection rate	90
4.6	Power curves	91
4.7	Histogram of Data 1 and fitted model, the red line shows the two-component fitted model, and the green line shows the three-component fitted model.	94
4.8	Histogram of Data 2 and fitted model, the red line shows the two-component fitted model, and the green line shows the three-component fitted model.	95

4.9	Histogram of Data 3 and fitted model, the red line shows the two-component fitted model, and the green line shows the three-component fitted model.	98
4.10	Histogram of Examples of Data 4 and fitted model, the red line shows the two-component fitted model, and the green line shows the three-component fitted model.	99

Chapter 1 Introduction

1.1 Finite Mixture Model

1.1.1 Definition of finite mixture model

To define a finite mixture model, we consider a random sample of size n : Y_1, \dots, Y_n , let $f(y; \theta_i) : \theta_i \in \Theta$ be a parametric family of probability density functions, then we suppose the density $f(y; \theta)$ of Y can be written in the form:

$$f(y; \theta) = \sum_{i=1}^g \pi_i f(y, \theta_i) \quad (1.1)$$

where $0 \leq \pi_i \leq 1$ and $\sum_i^g \pi_i = 1$. The $f(y, \theta_i)$ are densities which are called the *component densities of the mixture*, π_i are called the *mixing proportions*. Then $f(y; \theta)$ is called *g-component finite mixture density*. [McLachlan, 1994]

We consider g is a fixed number we have known in the formula (1.1). But when we deal with real data, we don't know the number of components and have to estimated g from data. For example, among the following models, (1.3) and (1.4) are equivalent, since (1.3) has a component that has a weight of zero while the last two-component of (1.4) can be combined to $Beta(1,1)$.

$$Beta(1, 1) \quad (1.2)$$

$$Beta(1, 1) + 0Beta(2, 1) \quad (1.3)$$

$$\frac{1}{3}Beta(1, 1) + \frac{1}{3}Beta(2, 1) + \frac{1}{3}Beta(1, 2) \quad (1.4)$$

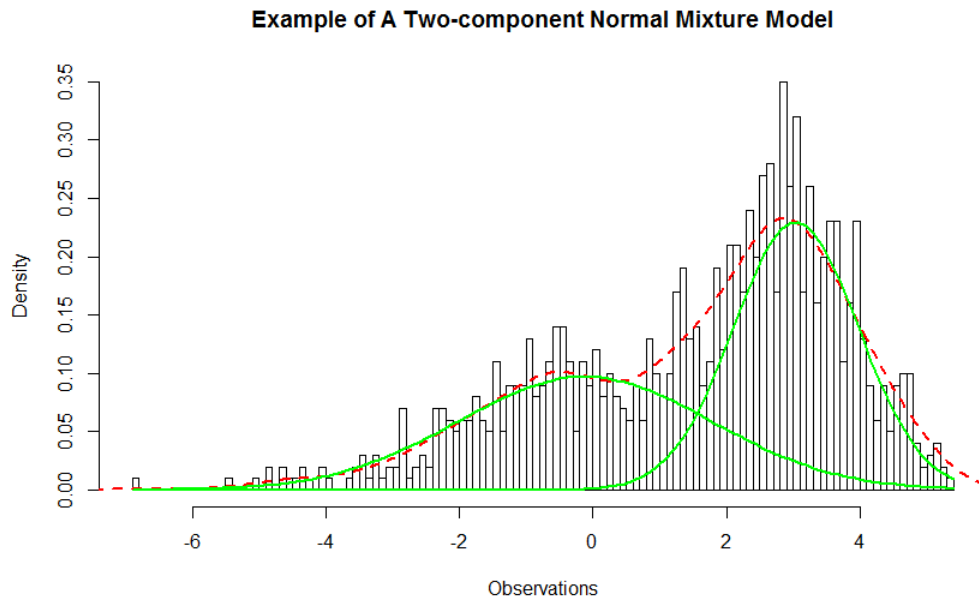


Figure 1.1: An example of two-component normal mixture model

1.1.2 Application of finite mixture model

The finite mixture model is an appealing strategy for dealing with a complicated distribution since it can precisely describe various shapes. For example, in figure 1.1, I show the histogram of a sample of size 1000 generated randomly from two normal distributions $N(3,1)$ and $N(0,2)$. A regular model such as Normal, Beta, or Gamma distribution can not describe the data accurately. In this case mixture model is a reasonable choice. I fit a two-component normal mixture model with R package "mixtools" [Benaglia et al., 2009]. On the other hand, although the non-parametric model is also widely used in this condition, we can get information on subpopulations and natural interpretation with the finite mixture model.

Because of the two main advantages we mentioned above, researchers use finite mixture models in many specific fields. For example, [Roeder, 1994] analyzed a data which contains 190 sodium-lithium counter-transport measurements. She uses a

graphic technique to determine the components number of the mixture model and get a three-component normal mixture with equal variance as a suitable model. [Chen et al., 2010] invented a new method, the expectation-maximization (EM) test, to analyze this data and conclude that a two-component normal mixture with unequal variance gives a better fit to it.

Another example is [Charnigo et al., 2010] paper tried to describe the birth weight distribution for a “population of white singleton infants born to heavily smoking mothers”. [Charnigo et al., 2010] use flexible information criterion(FLIC) to determined the number of components in the normal mixtures model. They finally conclude that a four-component normal mixture model is the most suitable model in describing the infants birth weight data.

1.2 Contaminated Density Model

In model (1.1), $\theta_1, \dots, \theta_g$ and π_1, \dots, π_g are often considered unknown, but in some models, such as “*contaminated density model*”, θ_1 is setting to be known if we have ‘suitable’ knowledge of the subjects while the other θ ’s are treated as unknown.

Beta contamination distributions are used to model the P-values in hypothesis testing of microarray experiments because beta distributions have wide range of different shape shapes distribution on the interval $[0,1]$, and Uniform(0,1) is a special case of beta distribution. [Allison et al., 2002]

[Dai and Charnigo, 2008] use omnibus test with the contaminated beta model (1.5) in gene filtration study. The omnibus test is highly recommended when people need to deal with large-scale hypothesis testing. That is because when the number of tests

is not very large, Type I error adjustments work well, but if there are thousands of testing, these adjustments will increase the false-negative rate. The omnibus test can overcome this difficulty.

$$(1 - \gamma)Beta(1, 1) + \gamma Beta(\alpha, \beta) \tag{1.5}$$

Where $\gamma \in [0,1]$ corresponds to the proportion of genes in the batch that are differentially expressed, and $1 - \gamma$ is the proportion of genes that are not differentially expressed. The notation is as in [Dai and Charnigo, 2008]

All P-values from thousands of tests could be regarded as a random sample from this Beta contamination model. The P-values for not differentially expressed genes are viewed as independently and identically sample from Uniform(0,1)(or Beta(1,1)), while the Beta distribution Beta(α, β) characterize the P-values of those differentially expressed genes.

They use the omnibus test

$$H_0 : (\alpha - 1, \beta - 1)\gamma = (0, 0) \tag{1.6}$$

$$H_1 : (\alpha - 1, \beta - 1)\gamma \neq (0, 0) \tag{1.7}$$

Obviously, under H_0 , the Beta contamination model above could be simplified to Uniform(0,1)(or Beta(1,1)), which implies there is little evidence to conclude there is a differential expression of those genes. Thus we could pick up the genes that with H_0 are rejected for further study.

An interesting thing is the number of contamination components of a contaminated

Beta model is ambiguous unless we place some constraints on the parameters. For example, (1.2) and (1.4) above can not be distinguished. We will have further discussion of this topic in detail in the following Chapters.

[Dai and Charnigo, 2010] also proposed a different approach to do the microarray data analysis. They use the contaminated normal model(1.8) to describe the distribution of Z statistics or transformed T statistics instead of using a contaminated beta model to model P-values.

$$(1 - \pi)Normal(0, \sigma^2) + \pi Normal(\mu, \sigma^2) \tag{1.8}$$

Then the corresponding omnibus test becomes

$$H_0 : \pi\mu = 0 \tag{1.9}$$

$$H_1 : \pi\mu \neq 0 \tag{1.10}$$

The notation is as in [Dai and Charnigo, 2010].

Similar to the Beta contamination model, let $\pi \in [0,1]$ be the proportion of genes in the batch that are differentially expressed. Also the Z statistics for the genes that are without expression alteration are $N(0; \sigma^2)$ for some $\sigma^2 > 0$, while the Z statistics of genes that are differentially expressed are $N(\mu; \sigma^2)$. σ is a nuisance parameter common to all components of a normal contaminated model and could be regarded as both known and unknown. As we may not need the information of σ^2 , setting it to be one is reasonable.

As contaminated Normal model can detect the direction of differential expression

since we can perform left-sided tests or right-sided tests with the contaminated Normal model to detect a batch of the under-expressed or over-expressed genes. Hence the Normal contamination model is more powerful than the Beta contamination model when the over-expression is more than the under-expression.

1.3 Zero-one-inflated beta model

As we mentioned in section 1.2, the beta density could describe multiple type of shapes between 0 to 1, so it is commonly used to describe the distribution of proportions. The problem is as beta distribution is continuous, we know the probability of any particular point is 0. If the proportion data contains a lot of 0's or 1's, we may consider a mixed continuous-discrete distribution to provide a better description to the data.[Young et al., 2022]

[Ospina and Ferrari, 2012] proposed the Zero-one-inflated beta (ZOIB) model, it is a mixed continuous-discrete model with three-component, where one component is a beta distribution, and the other two components are degenerate distributions at the values of 0 and 1.

According to [Ospina and Ferrari, 2012], that the probability density function could be defined as,

$$f(y; \pi_1, \pi_2, \alpha, \beta) = \begin{cases} \pi_1 & \text{if } y=0 \\ \pi_2 & \text{if } y=1 \\ (1 - \pi_1 - \pi_2)f(y; \alpha, \beta) & \text{if } y \in (0, 1) \end{cases} \quad (1.11)$$

where $f(y; \alpha, \beta)$ is the beta density, π_1 and π_2 are the probability mass at 0 and 1. The beta component represent the continuous proportions in the data, and the two degenerate distribution at 0 and 1 represent the zeros and ones in the proportion data.

[Ospina and Ferrari, 2012] take model selection criteria to estimate the parameter in their paper and [Wieczorek and Hawala, 2011] use a Bayesian approach to obtain the estimates.

1.4 Estimate number of components by Testing or Information Criteria

It is not easy work to give statistical inference of mixture modeling. People developed different kinds of approaches to estimate the number of mixture components, including hypothesis tests and information criteria.

Test

because we know LRT(likelihood ratio test) is a locally most powerful test, it usually is a reasonable choice. [Chen et al., 2001]

Suppose X_1, \dots, X_n be a random sample of size n from a two-component mixture model and the following formula is the ordinary log-likelihood function.

$$l_n(\pi, \theta_1, \theta_2) = \sum_{i=1}^n \log[(1 - \pi)f(X_i; \theta_1) + \pi f(X_i; \theta_2)], \quad (1.12)$$

where notation is as in [Chen et al., 2001].

[Dacunha-Castelle and Gassiat, 1999] showed that the LRT statistic will converges in law to $\sup_{\theta \in \Theta} (max(0; W(\theta)))^2$ under some condition, where $W(\theta)$ is a Gaussian pro-

cess. [Chen et al., 2001] point out if regularity conditions are violated in the mixture problem, then for testing homogeneity against a mixture alternative, the classical LRT statistic does not maintain the simple asymptotic structure.

So they introduced a new test: the MLRT (Modified likelihood ratio test). [Chen et al., 2001] proved the asymptotic properties of MLRT and mentioned it has similar power as LRT. They define a penalized log-likelihood function as follows:

$$pl(\pi, \theta_1, \theta_2) = \sum_{i=1}^n \log[(1 - \pi)f(X_i; \theta_1) + \pi f(X_i; \theta_2)] + C \log(4\pi(1 - \pi)) \quad (1.13)$$

Where $C \log(4\pi(1 - \pi))$ is the penalty term, notation is as in [Chen et al., 2001].

M_n is the MLR test statistic

$$M_n := 2pl_n^*(\hat{\gamma}, \hat{\theta}_1, \hat{\theta}_2) - 2pl_n^*\left(\frac{1}{2}, \hat{\theta}_0, \hat{\theta}_0\right) \quad (1.14)$$

We can see with adding penalty term, we could avoid the estimator of π going to 0 or 1. Then under basically the similar conditions as LRT, MLRT statistic converges in law to $\max(0; W(\theta_0)^2)$. Under the null hypothesis, test statistic M_n will converge in law to its limiting distribution $0.5\chi_0^2 + 0.5\chi_1^2$ (where χ_0 is a degenerate random variable at value 0). [Chen et al., 2001]

To solve this problem, [Charnigo and Sun, 2004] also introduce another method: D-test. When using D-test, we propose to have a fitted null model and a fitted alternative model, then measure the L2 distance between them. D-test statistics depend on parameter estimates instead of data itself, so this test has a greater advantage than MLRT when the data set given is not available, but parameter estimate is.

The D-test statistic is

$$d_n := \int [(1 - \hat{\alpha})f_{\hat{\sigma}}(x, \hat{\theta}_1) + \hat{\alpha}f_{\hat{\sigma}}(x, \hat{\theta}_2) - f_{\hat{\sigma}_0}(x, \hat{\theta}_0)]^2 dx \quad (1.15)$$

The notation is as [Charnigo and Sun, 2004].

Also, [Charnigo and Sun, 2010] showed asymptotic equivalences between the D-test and likelihood ratio-type test for homogeneity. For example, under the null hypothesis, as $n \rightarrow \infty$, $nd_n = C^*(\theta_0)M_n + o_p(1)$.

Most of the testing methods required regularity conditions, including the finiteness of Fisher information and the parameter space being compact. To solve the problem, [Li et al., 2009] proposed the EM-test, which does not need these assumptions by defining a penalized log-likelihood function similar to the MLRT, but the penalty part of it is different from MLRT.

$$pl(\pi, \theta_1, \theta_2) = \sum_{i=1}^n \log[(1 - \pi)f(X_i; \theta_1) + \pi f(X_i; \theta_2)] + C \log(1 - |1 - 2\pi|) \quad (1.16)$$

The EM test statistic is

$$En(\alpha_0) := 2pl_n^\dagger(\hat{\alpha}, \hat{\theta}_1, \hat{\theta}_2, \hat{\sigma}) - 2pl_n^\dagger\left(\frac{1}{2}, \hat{\theta}_0, \hat{\theta}_0, \hat{\sigma}_0\right) \quad (1.17)$$

Where $C \log(1 - |1 - 2\pi|)$ is the penalty term, and the notation is as in [Li et al., 2009].

This test has an advantage, before constructing the EM test statistics, it estimate parameters with fewer iterations. The EM test also has a simple limiting distribution: $E_n^{(k)} \rightarrow 0.5\chi_0^2 + 0.5\chi_1^2$ in distribution under the H_0 , note k is the number of iteration.

Testing the hypothesis of two-component versus more components with a finite mix-

ture model is also useful in applications. In some cases, researchers may be interested in determining the number of components needed to describe the data adequately and prefer less complex models for parsimony. Since there was no testing procedure for the hypothesis $g = 2$ versus $g \geq 3$, [Chen et al., 2004] considered testing for a finite mixture model with k components and the kernel distribution of the finite mixture model from a one-parameter family. They proposed a modified likelihood ratio statistic and showed the asymptotic null distribution.

They define a modified likelihood function

$$pl(\pi, \theta_1, \theta_2, \dots, \theta_g) = \sum_{i=1}^n \log[\pi_1 f(X_i; \theta_1) + \pi_2 f(X_i; \theta_2) + \dots + \pi_g f(X_i; \theta_g)] + C_g \sum_{j=1}^g \log(\pi_j) \quad (1.18)$$

Where $C_g \sum_{j=1}^g \log(\pi_j)$ is the penalty term, $\sum_{j=1}^g \pi_j = 1$, C_g is some constant, notation is as in [Chen et al., 2004].

They found if the kernel distribution satisfies some regularity conditions, the asymptotic limiting distributions of the modified LRT statistic R_n follow the mixture of χ^2 -distribution as

$$\left(\frac{1}{2} - \frac{\alpha}{2\pi}\right)\chi_0^2 + \frac{1}{2}\chi_1^2 + \frac{\alpha}{2\pi}\chi_2^2 \quad (1.19)$$

where $\alpha = \cos^{-1}(\rho)$, α is between 0 and π and depends on the parametric family under investigation. It could be estimated by MMLEs under null hypothesis.[Chen et al., 2004].

Information criateria

To determine the proper numbers of components for mixture models, some people use model selection criteria, for example, Akaike information criterion (AIC, [Akaike, 1974]) or Bayesian information criterion (BIC, [Schwarz, 1978]), other than testing. The AIC has a penalty term $2C$, and BIC has a penalty of $\log(n)C$, where C is the complexity of the model. In [Lahiri, 2001] book, he proved AIC is inconsistent. AIC has a tendency to overestimate the number of components, while BIC tends to favor models with fewer components because of the heavier penalty. [Keribin, 2000] developed a penalized likelihood estimator, and it is almost surely consistent, but the penalty of his estimator does not depend on data.

Later a novel information criterion, singular Bayesian information criterion (sBIC), was introduced by [Drton and Plummer, 2017]), which provides a Bayesian approach to studying singular model selection problems. Models having Fisher information matrices that are possibly singular and not invertible are known as singular models. For these models, it is not possible, if the Fisher-information matrix is singular, to approximate the log-likelihood function with a large sample quadratic approximation for BIC, according to Watanabe ([Watanabe, 2009]).

However, sBIC can handle this situation. On the one hand, for regular models, sBIC gives identical results compared to BIC while circumventing the Monte Carlo algorithm during the calculation. On the other hand, sBIC is proved to be consistent while maintaining the link between Bayesian methods and log-marginal likelihood in a normal circumstance, under which regular BIC cannot be applicable due to the fact that Fisher information matrices are not invertible because of singularity. It is worth mentioning that a regular BIC penalty is equal to sBIC counterpart, if not stronger.

1.5 EM Algorithm

In order to calculate maximum likelihood estimates, a general approach called the EM algorithm was used. [Dempster et al., 1977] presented the strategy in detail. EM algorithm computes the MLE iteratively. In every iteration, there are two steps; first step is called the expectation step(E step), and the maximization step(M step) is the second step. During the expectation step, we compute the expectation of the log-likelihood function based on the current parameters estimates; in the maximization step, we compute new parameter estimates via maximizing the expectation of log-likelihood we obtained in the previous E step. Then the new parameter estimates are used to compute the log-likelihood in the next E step. We repeat the E step and M step until the process converge.

In this dissertation, we use the EM algorithm to estimate the parameters for mixture models that are unknown. The EM algorithm has advantages in mixture model parameter estimation: firstly, while circumventing the calculation of the numerical solutions for high-dimensional optimization problems, which are difficult and computational expensive, it can approximate the maximum likelihood estimation (MLE); secondly, according to Dempster ([Dempster et al., 1977]), if we assume that the number of components is unknown in the model, then EM algorithm is more interpretable when dealing with incomplete data.

Here is how we apply the EM algorithm to the mixture model problems. Notation is in [Qi, 2016] dissertation:

Suppose we have X_1, X_2, \dots, X_n be iid random variables from a finite mixture

model with g components.

$$\sum_{j=1}^g \pi_j f(x|\theta_j) \text{ where } \pi_j \in [0, 1] \text{ and } \sum_{j=1}^g \pi_j = 1 \quad (1.20)$$

Let w_{ij} =I [item i belongs to the j^{th} component], then we could express the complete data log-likelihood function as the following form

$$l(\boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{i=1}^n \sum_{j=1}^g w_{ij} [\log \pi_j + \log f(x_i|\theta_j)] \quad (1.21)$$

Then start with initial values and perform the E step and M step in each iteration.

1.5.1 E-step

Put

$$Q(\pi, \theta | \pi^{(t)}, \theta^{(t)}) = \mathbf{E}[l(\boldsymbol{\pi}, \boldsymbol{\theta}) | \boldsymbol{\pi}^{(t)}, \boldsymbol{\theta}^{(t)}] \quad (1.22)$$

Let

$$w_{ij} = \frac{\pi_j f(x_i^{(t)} | \theta_j^{(t)})}{\sum_g \pi_g f(x_i^{(t)} | \theta_g^{(t)})} \quad (1.23)$$

Then equation 1.21 becomes

$$\sum_{i=1}^n \sum_{j=1}^g w_{ij}^{(t)} \log \pi_j + \sum_{i=1}^n \sum_{j=1}^g w_{ij}^{(t)} f(x_i|\theta_j) \quad (1.24)$$

1.5.2 M-step

Maximize the function 1.23, we obtained the estimates after next iteration: $\pi^{(t+1)}, \boldsymbol{\theta}^{(t+1)}$.

Then we need to update function 1.21 with the new estimates, and repeat the iteration until the likelihood function converges.

$$|l(\pi^{(t+1)}, \theta^{(t+1)}) - l(\pi^{(t)}, \theta^{(t)})| < \epsilon \quad (1.25)$$

for some small $\epsilon > 0$.

When we use EM algorithm, the choice of initial values deserve our attention. Choice of initial values is essential because it strongly affect the convergence speed of the EM algorithm and whether it could reach the global optima. Various researchers have studied and discussed the choice of initial values of the EM algorithm for finite mixture models.

[Karlis and Xekalaki, 2003] review and compare several methods for choosing the initial values of EM algorithm. [Lahiri, 2001] mentioned that random initial values are the simplest choice. For example, reseachers can generate initial values of the parameter from uniform distributions and generate the initial values of the mixing proportions from uniform(0,1). Researchers could also consider choosing initial values by computing likelihood function. That means one can randomly generate many initial value sets and calculate the log-likelihood for each set, then choose initial values by selecting several "best" sets with the largest log-likelihood. [Karlis and Xekalaki, 2003] also introduced the grid search method developed by [Laird, 1978] for setting the initial values; and [Böhning et al., 1994] modified grid search method on big parameter space. [Furman and Lindsay, 1994] proposed using the estimates from the moment method as the initial values.

Copyright© Ya Qi, 2022.

<https://orcid.org/0000-0001-6656-7008>

Chapter 2 Two-component Beta Mixture Model with Constraints

2.1 Introduction

As we mentioned in Chapter 1, the beta mixture model could describe multiple shapes distribution on the interval $[0, 1]$; thus, in micro-array data analysis and large-scale hypothesis testing, we could use a beta mixture distribution to model the p-values from many tests. ([Allison et al., 2002]).

We also introduced the modified loglikelihood ratio test(MLRT) for homogeneity in mixture models with a general parametric kernel distribution family and its null limiting distribution in Chapter 1. ([Chen et al., 2001]). Additionally, [Dai and Charnigo, 2008] described an omnibus test with the Contaminated Beta model(CB model) to do gene filtration.

Figure 2.1 shows possible shapes of the two-component CB model densities. If we do the microarray analysis, the p-values for not differentially expressed genes could be described as a sample from Beta(1,1). If some of the genes are differentially expressed, the p-values could be modeled as a Contamination model with two-components Beta(1,1) and Beta(α, β)[Dai and Charnigo, 2008]. When we observed the real gene microarray data, we found that if the null hypothesis is false, the distribution of the p-values is right-skewed and concentrated to zero. It looks that the top right contaminated beta model in figure 2.1 could describe the distribution of P-values suitably for most microarray data. Thus, by designing a testing procedure with constraints to put the mode to the left end, the test would be more sensitive to the micro-array data. We expect by adding the constraints $0 < \alpha \leq 1 \leq \beta$ to this two-component CB model, a more precise alternative hypothesis would give a more

powerful test.

Consider the two-component Beta contamination model:

$$(1 - \gamma)Beta(1, 1) + \gamma Beta(\alpha, \beta) \quad (2.1)$$

The notations are same with the chapter 1.

Define a penalized log-likelihood function of Beta contamination model as follows:

$$Pl(\pi, \alpha, \beta) = \sum_{i=1}^n \log[(1 - \gamma)f(X_i; \alpha_0, \beta_0) + \pi f(X_i; \alpha, \beta)] + C \log(4\gamma(1 - \gamma)) \quad (2.2)$$

Where $(\alpha_0, \beta_0) = (1, 1)$, $C \log(4\gamma(1 - \gamma))$ is the penalty term, $\gamma \in (0, 1)$, $\alpha \in (0, 1]$, $\beta \in [1, \infty)$. C is some constant used to control the level of penalization.

2.2 Estimate the MMLEs with constraints

We need to maximize the penalized log-likelihood function above to obtain the MMLEs with constraints. As we have constraints on two parameters α and β when maximizing the log-likelihood function, obtaining the MMLEs becomes a box-constraint optimization problem.

As mentioned in chapter 1, the EM approach is a commonly used method to obtain the maximum-likelihood estimates for the mixture models. It has two steps for each iteration: E(expectation) step and M(Maximum) step. As we added constraints to the Beta contamination model, we need to do a constrained optimization during the M step.

The BFGS algorithm is a commonly used strategy when considering nonlinear-optimization

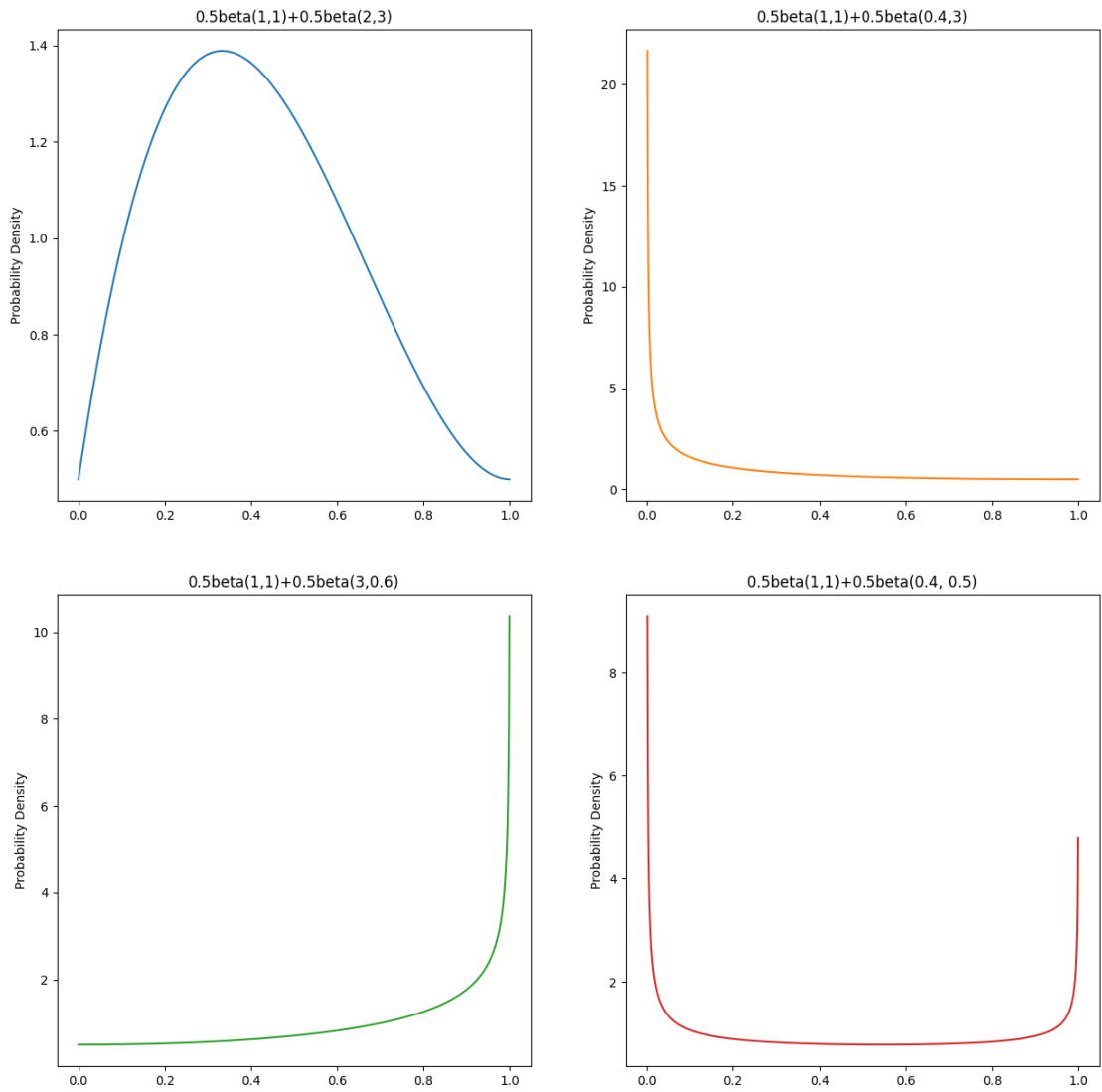


Figure 2.1: Basic shape of contamination Beta mixture models (modified from [Dai and Charnigo, 2008])

problems. BFGS is a quasi-Newton method. ([Fletcher, 2000]). Richard H. Byrd, Pei-huang Lu, Jorge Nocedal, Ciyou Zhu modified the BFGS algorithm and developed the L-BFGS-B algorithm; it could solve the box-constrained optimization problem (The variable we need to estimate have constraints with lower and upper bounds [Byrd et al., 1995]). "It is based on the gradient projection method and uses a limited memory BFGS matrix to approximate the Hessian of the objective function".[Byrd et al., 1995] In R, the L-BFGS-B algorithm is implemented as an option of the base function `optim()`.

We use penalty coefficient $C = 10$ according to [Dai and Charnigo, 2008]. [Chen et al., 2001] detailed discussion of the choice of penalty coefficient in section 2.1. We set upper bound of both parameter α and β as 20 for the MMLEs calculation without constraints and set upper bound of α and lower bound of β as 1 , upper bound of β is 20 for the MMLEs calculation with constraints. To increase the probability to obtain the global optimization of the likelihood function, we choose 3 sets of 'best' random initial values via the method we mentioned in chapter 1.

2.3 The homogeneity hypothesis testing

[Dai and Charnigo, 2008] use a reparametrization method to obtain the limiting distribution of the test statistic M_n under null. M_n is the MLRT statistic

$$M_n := 2Pl_n(\hat{\gamma}, \hat{\alpha}, \hat{\beta}) - 2Pl_n\left(\frac{1}{2}, \alpha_0, \beta_0\right) \quad (2.3)$$

where Pl_n is the penalized log-likelihood function.

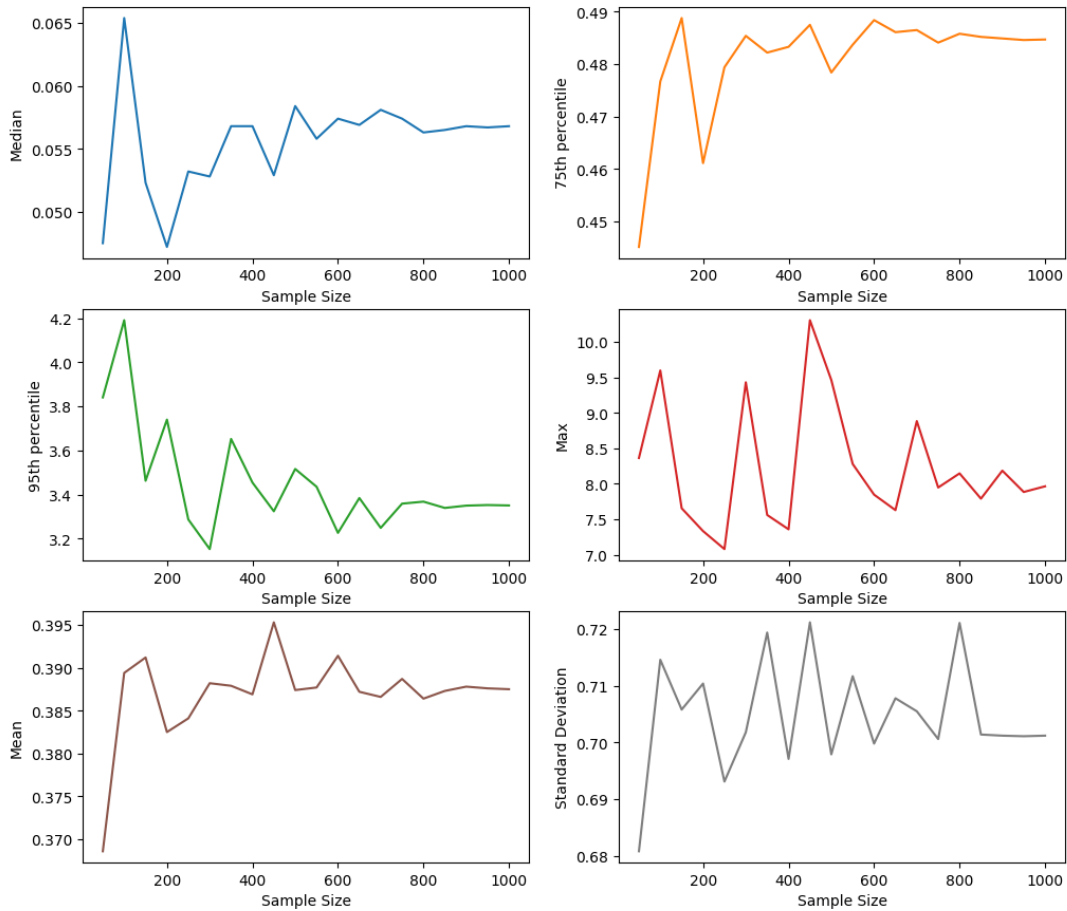


Figure 2.2: The Characteristics of MLRT vs simple size

They proved that if the parameter space is compact and (α_0, β_0) is belongs to its interior, then under the null, $M_n \xrightarrow{d} \chi_2^2$.

If we add the constraints $\alpha \in (0, 1]$, $\beta \in [1, \infty)$, we don't know the asymptotic null limiting distribution of the MLRT. Is it still have a χ^2 -type null limiting distribution?

Table 2.1: The Characteristic of the simulated MLRT with constraints(n=2000)

Characteristic	mean	variance	Q1	median	Q3	95th	max
value	0.38754	0.70121	0	0.05682	0.04872	3.35135	8.39563

2.3.1 The null limiting distribution

First, we considered doing some simulation study to check the null limiting distribution with the two-component beta contamination model was added constraints.

For each of several sample size (n=50, n=100, n=150, \dots , n=1000), I generate 5000 data sets from Beta(1,1) and calculate the modified LRT statistics. Then I get characteristics such as minimum, median, maximum, mean, standard deviation, and some quantiles of the modified LRT statistics. Figure 2.2 shows the trend of some characteristics changing as the sample size increases. As the minimum and 25th quantile are all zeros, I didn't include their plots in the figure.

We can see in Figure 2.2 that as the sample size increase, the characteristic becomes more and more consistent. Table 2.1 shows the characteristic.

Then I generate 5000 data sets from null distribution Beta(1,1) for sample size=2000 and calculate the MLRT for each data set. Figure 2.3 show the histogram of MLRT. When we observe the shape of the histogram, the possibility of it still having a χ^2 type limiting distribution could not be ruled out.

Then I assume the null limiting distribution of MLRT still has some mixture of χ^2 distributions and try the moment method to determine the weights based on the simulation results we get above.

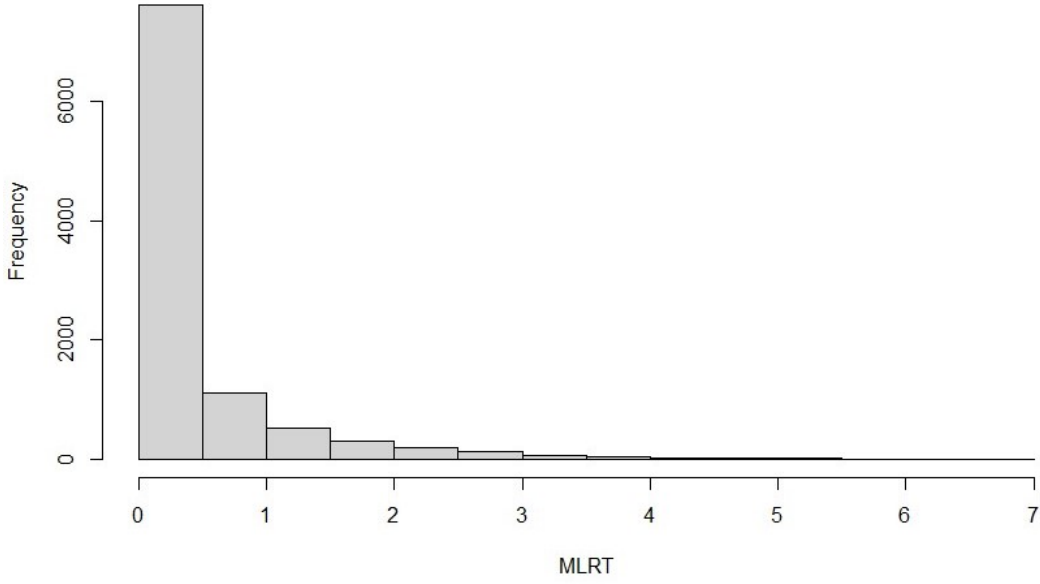


Figure 2.3: Histogram of simulated MLRT with constraints (n=2000)

For example, there is no harm to suppose $a \in (0, 1)$, $b \in (0, 1)$, we have

$$M_n \sim (1 - a - b)\chi_0^2 + a\chi_1^2 + b\chi_2^2 \quad (2.4)$$

According to the properties of χ^2 distribution, we get

$$E[M_n] = a + 2b \quad (2.5)$$

$$Var[M_n] = 2a^2 + 4b^2$$

Next, we have two equations by plugging in the estimates of mean 0.38754 and variance 0.70121 we get from the simulation above, then solving equations to get the estimates of a and b. we have $a = -0.3184$, $b = 0.3530$ or $a = 0.5768$, $b = -0.0946$. As a and b should be non-negative, so the two sets of solutions are invalid.

Otherwise, let's suppose

$$M_n \sim (1 - c)\chi_0^2 + c\chi_1^2 \quad (2.6)$$

we get

$$\begin{aligned} E[M_n] &= c \\ Var[M_n] &= 2c^2 \end{aligned} \tag{2.7}$$

If plugging in the simulated mean and variance, we could not find a real number solution of c to satisfy both equations.

From the results above, the limiting distribution of MLRT with constraints may not be a χ^2 mixture model. Or it may still be a χ^2 -type model, but the structure is very complicated to derive. As the null limiting distribution is not easy to get and use, using the actual critical value we obtained from the simulation is advisable.

2.4 Simulation study

2.4.1 Actual rejection rate

The next simulation studies the actual rejection rate under H_0 . We use critical point 3.35 from the simulation when the constraints are added. It is possible to use different critical value for different sample size, I will study this later in next subsection. The critical value we used for the alternative model without constraints is from the asymptotic theory, which is 5.99.

For sample size $n=50, n=100, \dots, n=500$, we generate 5000 data sets from the null distribution $Beta(1,1)$ and calculate the number of rejected null hypotheses out of 5000 based on the two different Beta contamination models (with or without constraints) and critical points. The results are shown in Figure 2.4. The nominal rejection rate is 0.05.

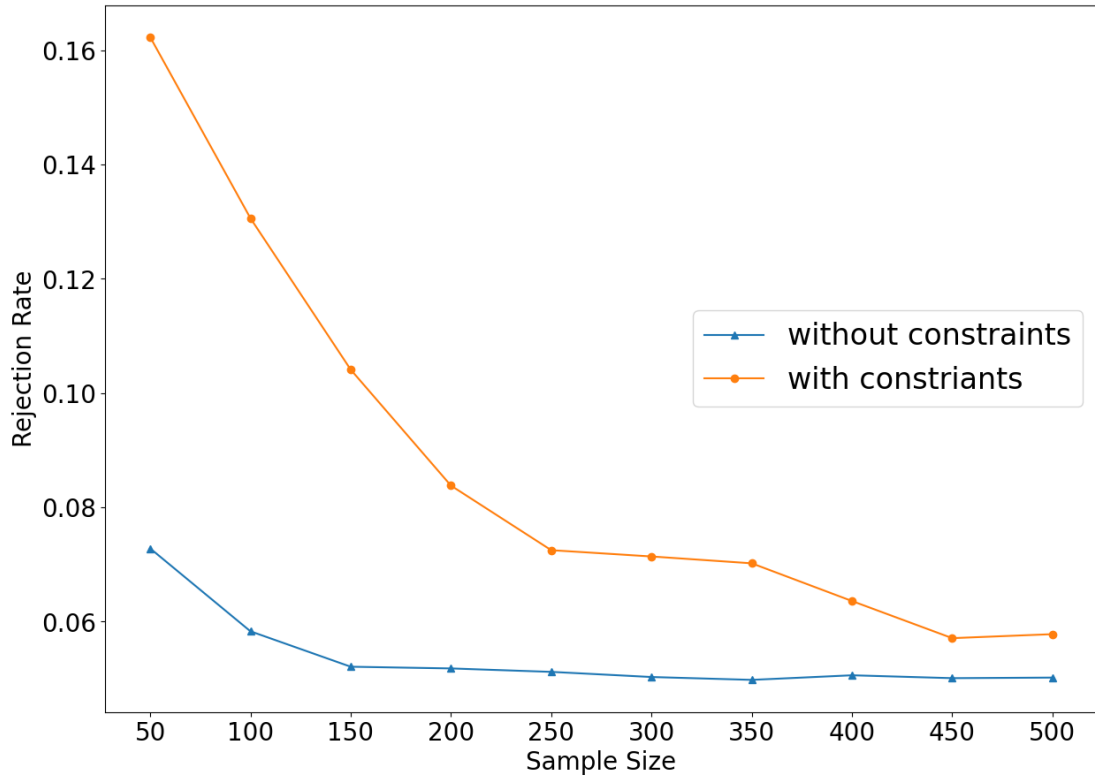


Figure 2.4: Actual rejection rates vs sample size

From Figure 2.4, we can see the actual rejection rates for the MLRT with constraints and without constraints become closer to the nominal rejection rate as the sample size increase. Even with small sample size, the actual rejection rate of MLRT without constraints is still not very far from 0.05. For the MLRT with constraints, we need a sample size larger than 200 to close to 0.05. When the sample size is smaller than 250, we may consider using an actual critical value.

Table 2.2: $0.75\text{Beta}(1,1)+0.25\text{Beta}(0.7,2)$

sample size	without constraints	with constraints	K-S test
50	0.368	0.335	0.281
100	0.592	0.618	0.544
150	0.74	0.765	0.705
200	0.817	0.852	0.793
250	0.884	0.922	0.872
300	0.91	0.945	0.907
350	0.932	0.974	0.924
400	0.958	0.988	0.948
450	0.978	0.989	0.957
500	0.983	0.996	0.964

2.4.2 Power

The second simulation study in this section, we are interested in the power under H_1 .

We generate 5000 data sets from

$$\begin{aligned}
 (I) & 0.75\text{Beta}(1, 1) + 0.25\text{Beta}(0.7, 2) \\
 (II) & 0.9\text{Beta}(1, 1) + 0.1\text{Beta}(0.5, 1.5) \\
 (III) & 0.95\text{Beta}(1, 1) + 0.05\text{Beta}(0.5, 1.5) \\
 (IV) & 0.95\text{Beta}(1, 1) + 0.05\text{Beta}(0.6, 3)
 \end{aligned} \tag{2.8}$$

We use critical point 3.35 when the constraints are added. The critical value we used without constraints is 5.99. As shown in Figure 2.5, the MLRT with constraints yields slightly better power than the MLRT without constraints. The performance of two MLRTs are better than the Kolmogorov-Smirnov test in all the four competitions of power, especially when the contamination fraction is insubstantial.

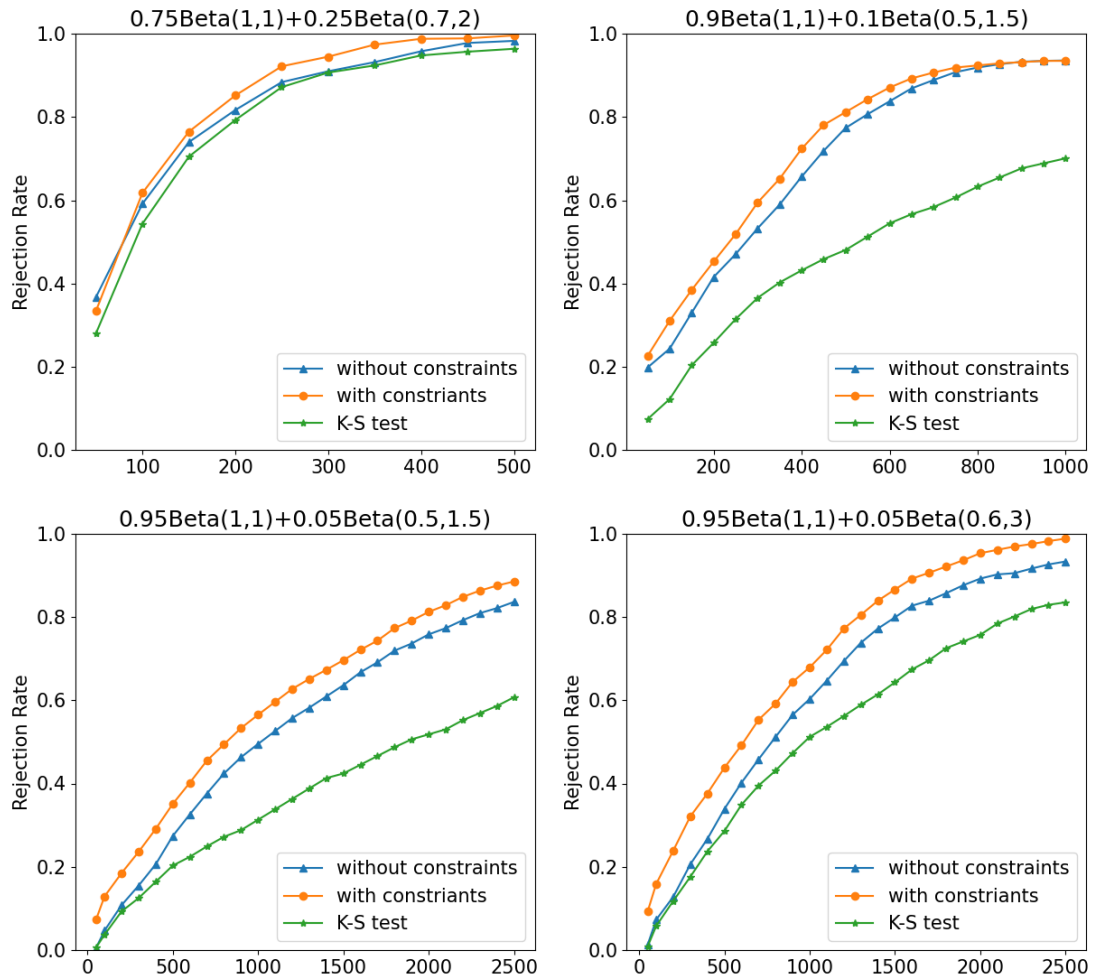


Figure 2.5: Power curve vs sample size

Table 2.3: $0.9\text{Beta}(1,1)+0.1\text{Beta}(0.5,1.5)$

sample size	without constraints	with constraints	K-S test
50	0.199	0.227	0.075
100	0.244	0.311	0.123
150	0.33	0.385	0.204
200	0.416	0.453	0.258
250	0.471	0.519	0.315
300	0.533	0.595	0.366
350	0.59	0.652	0.403
400	0.657	0.724	0.432
450	0.719	0.781	0.459
500	0.774	0.812	0.481
550	0.807	0.843	0.513
600	0.838	0.871	0.545
650	0.869	0.893	0.567
700	0.889	0.907	0.584
750	0.908	0.919	0.607
800	0.919	0.924	0.633
850	0.927	0.929	0.655
900	0.933	0.932	0.677
950	0.935	0.935	0.689
1000	0.936	0.935	0.701

Then we generate 5000 samples from

$$\begin{aligned}
 (V) & 0.5\text{Beta}(1, 1) + 0.5\text{Beta}(0.2, 0.9) \\
 (VI) & 0.5\text{Beta}(1, 1) + 0.5\text{Beta}(1.1, 4) \\
 (VII) & 0.5\text{Beta}(1, 1) + 0.5\text{Beta}(2, 3) \\
 (VIII) & 0.75\text{Beta}(1, 1) + 0.25\text{Beta}(2, 3) \\
 (IX) & 0.9\text{Beta}(1, 1) + 0.1\text{Beta}(1.5, 0.5) \\
 (X) & 0.95\text{Beta}(1, 1) + 0.05\text{Beta}(0.6, 0.7)
 \end{aligned} \tag{2.9}$$

The six scenarios above represent the Beta contamination densities with different shapes and contamination violate the constraints $\alpha \in (0, 1]$ and $\beta \in [1, \infty)$. Like we did in the previous simulation, we compute the number of rejected null hypothesis out of 5000 based on the MLRT with two different Beta contamination models (with

Table 2.4: $0.95\text{Beta}(1,1)+0.05\text{Beta}(0.5,1.5)$

sample size	without constraints	with constraints	K-S test
50	0.005	0.073	0.006
100	0.047	0.129	0.037
200	0.108	0.185	0.093
300	0.155	0.236	0.125
400	0.206	0.291	0.164
500	0.274	0.351	0.203
600	0.326	0.402	0.224
700	0.376	0.455	0.249
800	0.424	0.494	0.272
900	0.463	0.533	0.288
1000	0.495	0.565	0.313
1100	0.526	0.596	0.338
1200	0.557	0.627	0.363
1300	0.582	0.651	0.388
1400	0.609	0.673	0.413
1500	0.636	0.696	0.424
1600	0.667	0.721	0.445
1700	0.691	0.743	0.466
1800	0.719	0.773	0.487
1900	0.736	0.791	0.506
2000	0.758	0.812	0.518
2100	0.773	0.828	0.53
2200	0.792	0.848	0.552
2300	0.809	0.863	0.569
2400	0.821	0.875	0.586
2500	0.836	0.885	0.607

or without constraints) and critical points, respectively. The power curves are shown in Figure 2.6 and 2.7

As shown in Figure 2.6, the parameter of scenario (V) and (VI) violate the constraints mildly: for (V), $\beta=0.9$ is 0.1 small than 1 and α in (VI) is 0.1 larger than 1. The MLRT with the constraints still gives a good power curve in (V) and (VI). Figure 2.7 shows that when we generate data sets from the Beta contamination model with constraints that are violated heavily, we can see the power of MLRT with constraints becomes quite low; as the sample size increase, the power slightly decrease. On the

Table 2.5: $0.95\text{Beta}(1,1)+0.05\text{Beta}(0.6,3)$

sample size	without constraints	with constraints	K-S test
50	0.013	0.093	0.009
100	0.073	0.159	0.058
200	0.128	0.239	0.118
300	0.206	0.321	0.176
400	0.267	0.376	0.237
500	0.339	0.438	0.286
600	0.402	0.492	0.349
700	0.458	0.553	0.395
800	0.512	0.592	0.431
900	0.565	0.644	0.473
1000	0.603	0.679	0.512
1100	0.647	0.721	0.536
1200	0.694	0.772	0.562
1300	0.738	0.805	0.589
1400	0.772	0.839	0.614
1500	0.799	0.866	0.643
1600	0.827	0.892	0.674
1700	0.839	0.906	0.696
1800	0.857	0.921	0.725
1900	0.876	0.936	0.741
2000	0.892	0.953	0.757
2100	0.902	0.961	0.784
2200	0.905	0.969	0.801
2300	0.916	0.975	0.819
2400	0.926	0.982	0.829
2500	0.933	0.988	0.835

other hand, the MLRT without constraints still has good performance.

Figure 2.8 shows an example; when we generate one sample ($n=500$) from each scenario, calculate the MMLEs with and without constraints, respectively, and plot the true parameter value and two MMLEs for each scenario, we can see when we use MLRT with constraints. However, true parameter is violated the constraints, our MMLEs still are bounded within ($0 \leq \alpha \leq 1 \leq \beta$). If the violation is mild, the estimates are quite close to the true value, but if the violation is severe, it will make our MMLEs get closer and closer to $\alpha =1$ and $\beta =1$ as sample size increase. Although the

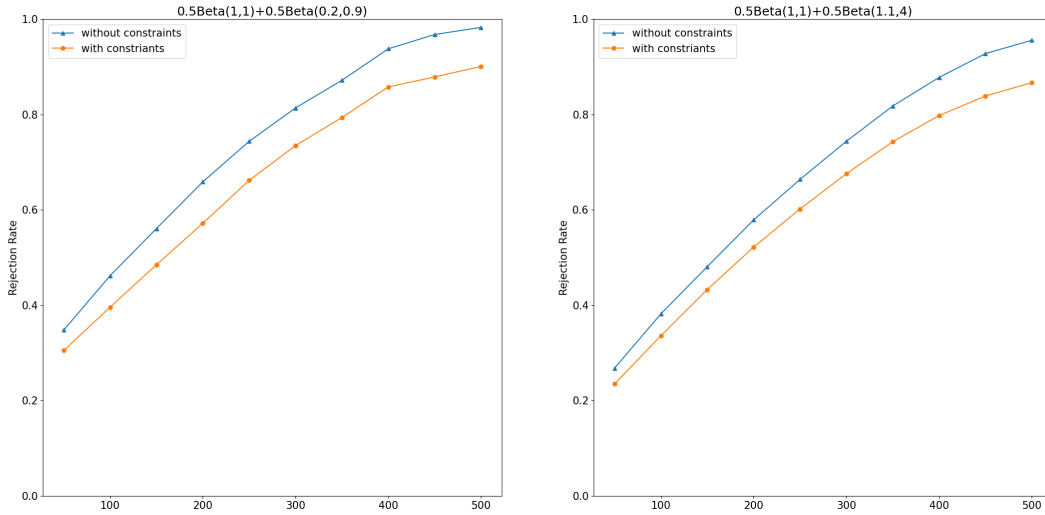


Figure 2.6: Power curves vs sample size of (V)-(VI)

true model is two-component, the MLRT is too small to reject the null. If we check the density of scenario (V) – (X), which violates the constraints, we could find when the violation is mild (scenario (V) and (VI)), the density is still close to the shape we mentioned at the beginning of the chapter: concentrated to 0 and right-skewed, but the shapes of (VII) – (X) is not.

To summarize, the simulation study demonstrated that when the p-values are concentrated near 0 and right-skewed, it may have the advantage of using the MLRT with constraints in most cases, which is usually true with most microarray analysis. But when the p-value distribution is not in this shape, the ordinary MLRT without constraints is preferred. Before we use the constrained model, we must check the distribution of real data. The goodness of fit test is not advisable to compare to the MLRT because its alternative space is larger compared to MLRT: its alternative hypothesis is the p-values follow a model other than Beta(1,1).

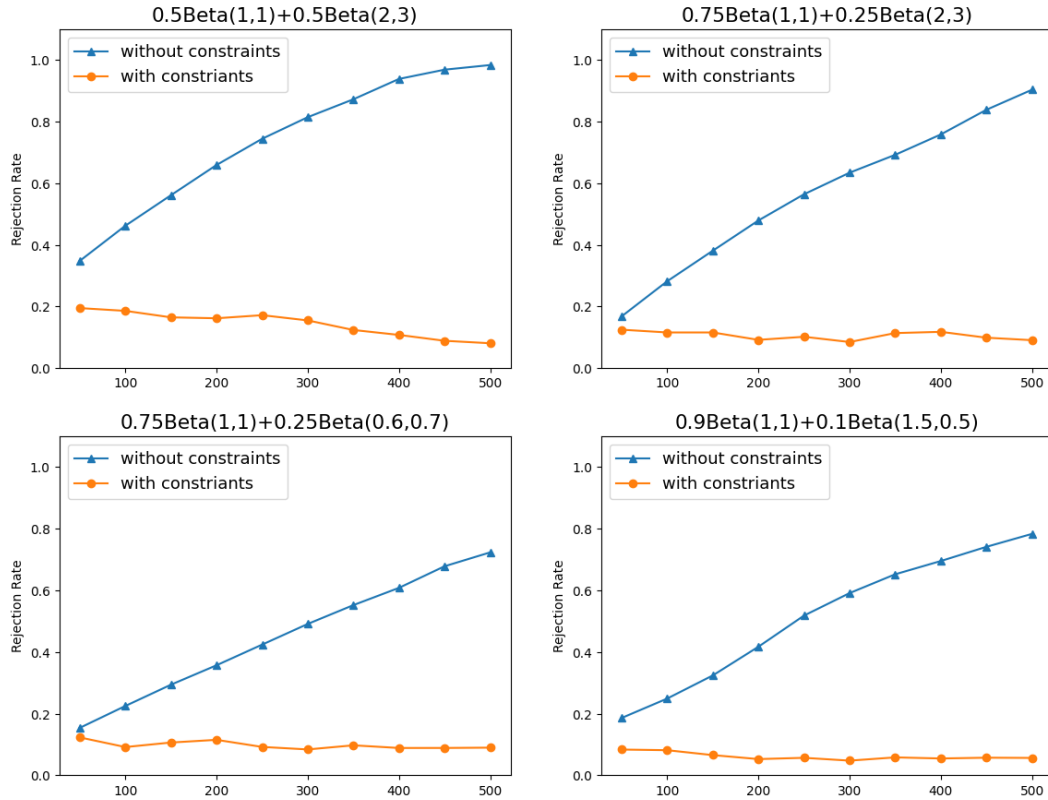


Figure 2.7: Power curves vs sample size of (VII)-(X)

2.4.3 Interpolation with Small sample

As we showed in the first part of the simulation study, we may consider using an actual critical value when the sample size is smaller than 250. We simulated the critical value of sample size $n=50, 60, 70, \dots, 240, 250$ by bootstrapping respectively. I interpolated between points of sample size and critical value. The critical value is a decreasing function of sample size; with a bunch of points, we fit a linear model to determine the critical value in terms of sample size and got

$$\hat{C}V = -0.00768 \times n + 5.1173 \quad (2.10)$$

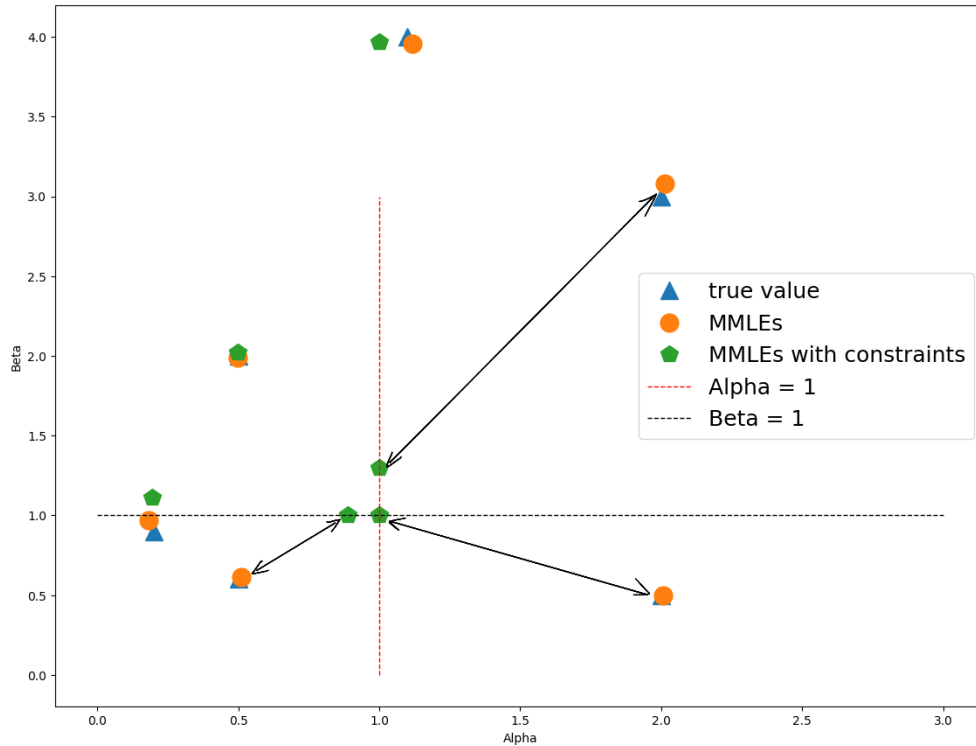


Figure 2.8: Compare MMLEs and true parameter

Where CV is critical value and n is sample size.

Then we don't need to estimate the actual critical value by bootstrapping every time we deal with a small sample. And the power curves in Figure 2.11 show, when we use the critical value estimated by the linear model, the power is slightly lower than the MLRT with constraints (use critical value 3.35) but still better than the MLRT without constraints (use critical value 5.99). When we deal with a small sample, we could use the equation to estimate critical value by sample size n and avoid repetitive bootstrapping. In this part, we choose a simple linear model to estimate the critical value, people could also choose different model to do the interpolation.

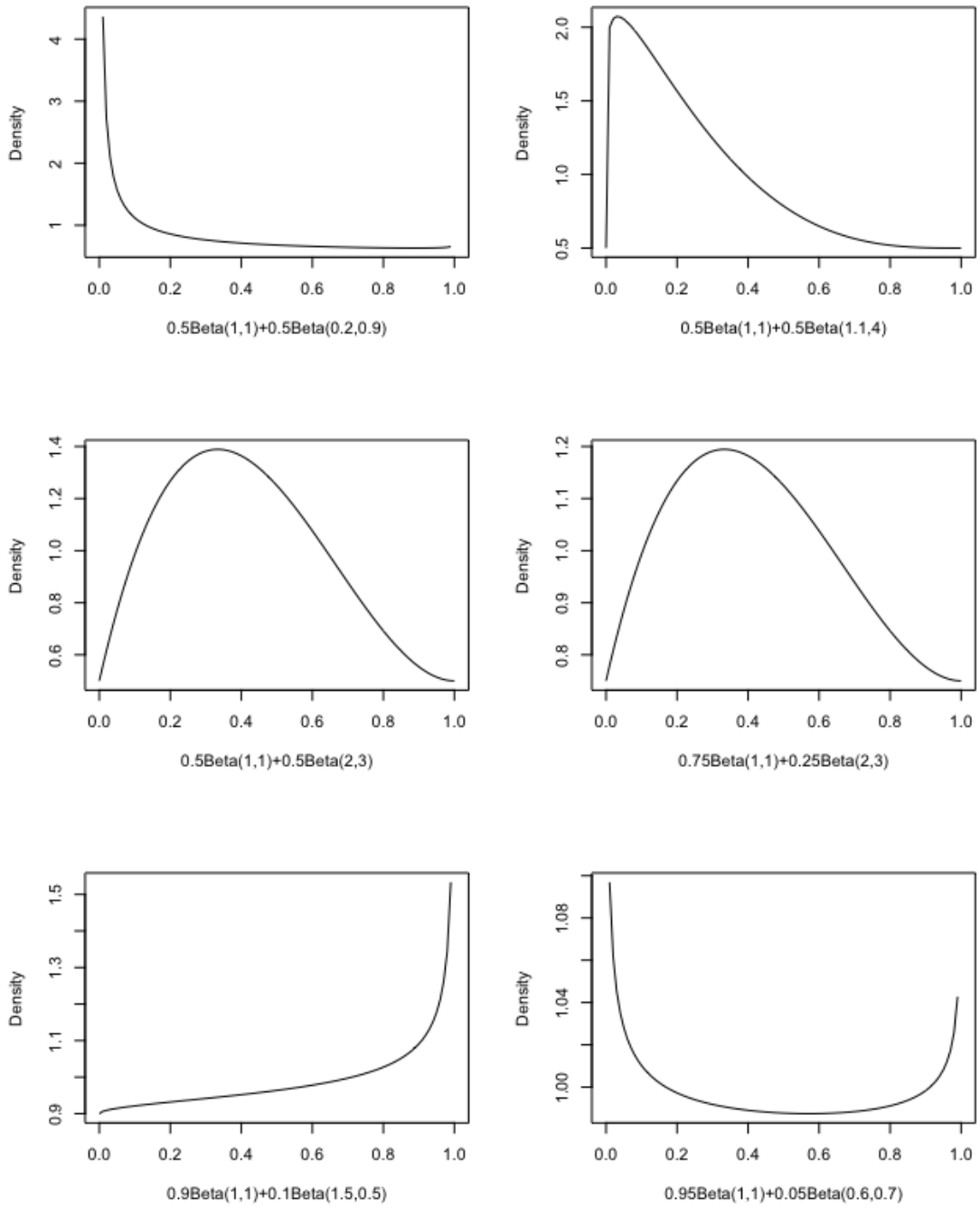


Figure 2.9: Shape of scenario (V) – (X)

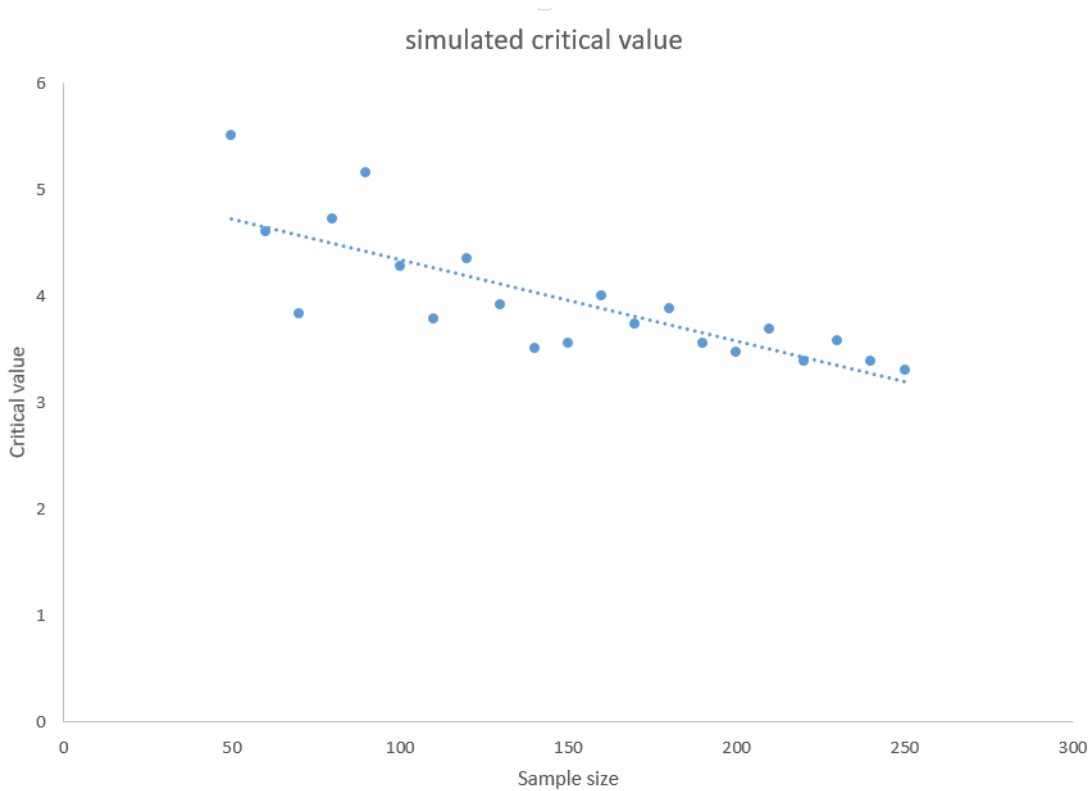


Figure 2.10: Estimate critical value when sample size is small

2.5 Real data application

2.5.1 introduction to the data

[Naumova et al., 2021] reported data on the ‘systematic genome-wide DNA methylation alternation in blood cells of toddlers with Down Syndrome’. There are 34 children whose age are from 0.5 to 4.5 years participated this study. 17 of the participant are children with the Down syndrome, 6 girls and 11 boys in the group; the other 17 participants are normally developing children, 7 girls and 10 boys are in this group. The mean age of Down syndrome group children are 33.88 month with standard deviation of 16.22. For the normally developing children group, their mean age is 33.35 month with standard deviation of 11.28. Participant’s age and gender proportion was not significantly different between the groups. Researchers also claim

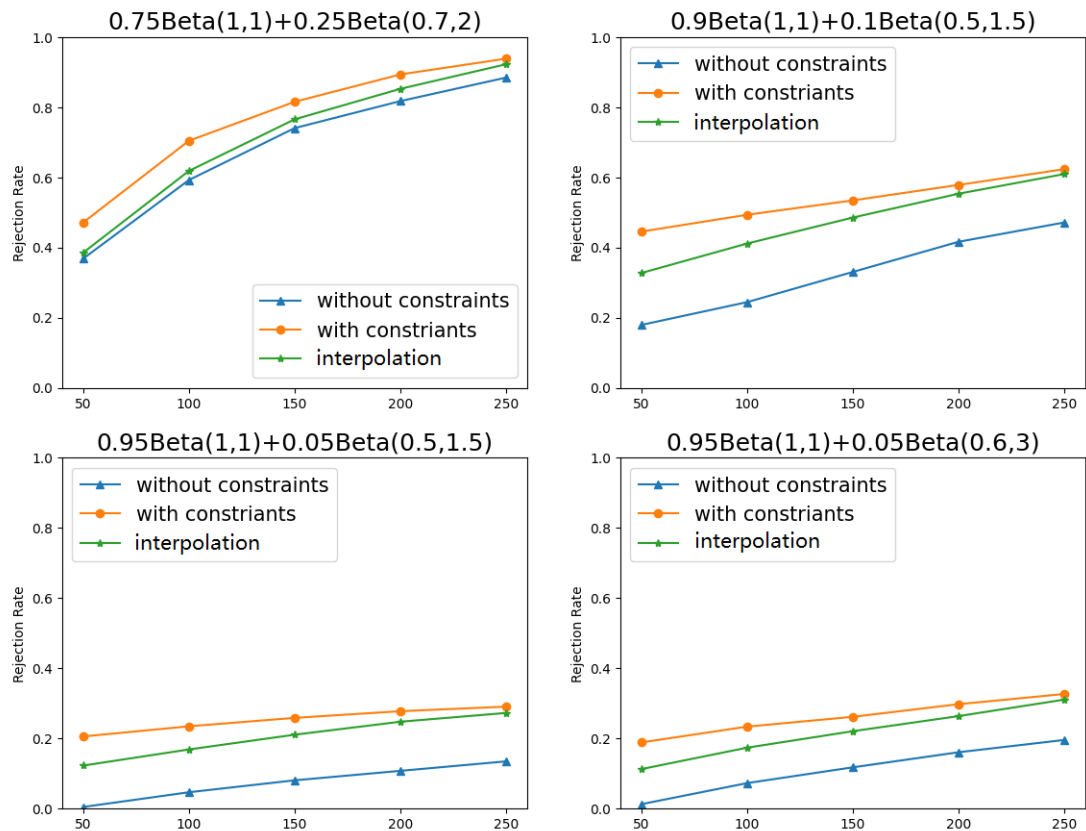


Figure 2.11: Power curves when sample size is small

that the children from both groups shared the same living environment and received the same care. The data is available for download on the following website.

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE174555>.

Data 1: The microarray contains genome-wide probes for 485,577 methylation sites. There are 461,258 methylation left after dropped missing values. Then the researchers perform T-tests to compare the gene expression level of all remaining methylation between the normally developed children group and the Down Syndrome children group. Finally, they obtained numerous p-values, one for each methylation site.[Naumova

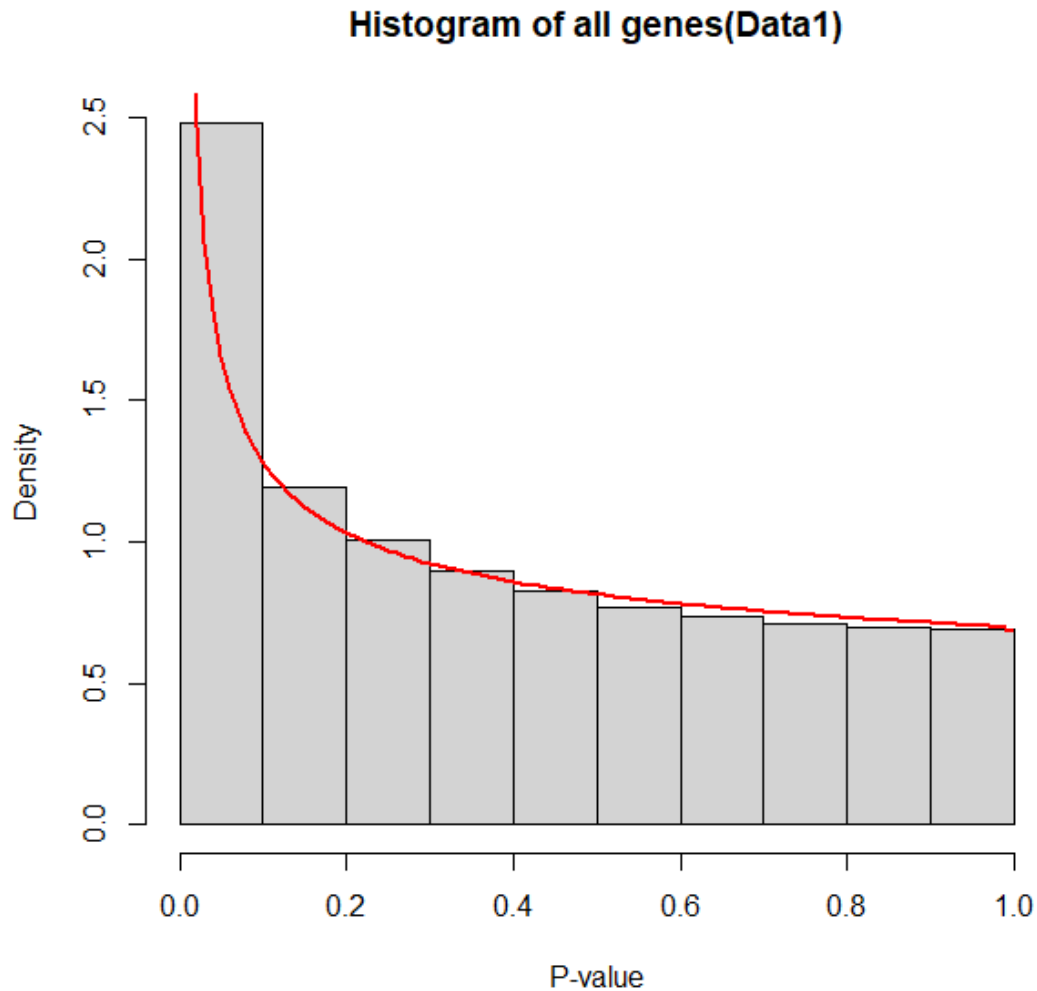


Figure 2.12: Histogram of all p-values, n=461258

et al., 2021]

Data 2: As we are interested in the Down Syndrome, we select the p-values of methylation site which located on Chromosome 21 according to the "Illumina Human Methylation 450k". The sample size is 4205. The document 'Illumina Human Methylation 450k' is available on website below:

Infinium HumanMethylation450K v1.2 product files

Data 3: Then we eliminate the methylation site containing DS-associated(Down

Histogram of p-values of CHR 21(Data2)

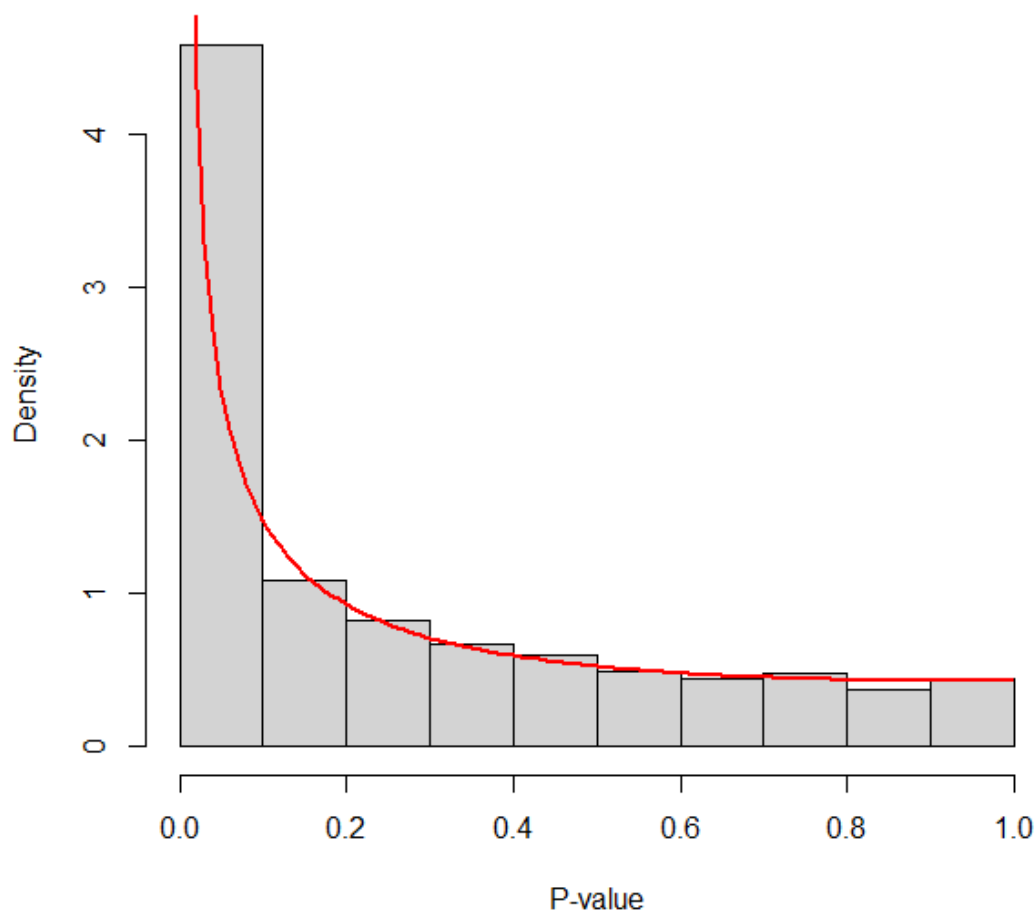


Figure 2.13: Histogram of p-values on chromosome 21

Syndrome-associated) differentially CpG sites according to the list from [Naumova et al., 2021] and 'Illumina Human Methylation 450k', the 452477 gene were left in the data 3. CpG site is regions on which a "cytosine nucleotide next to a guanine nucleotide, and lined by a phosphate group" [Jabbari and Bernardi, 2004]

Data 4: Finally, we random select 200 samples without replacement with sample size $n = 100$ from the Data 1.

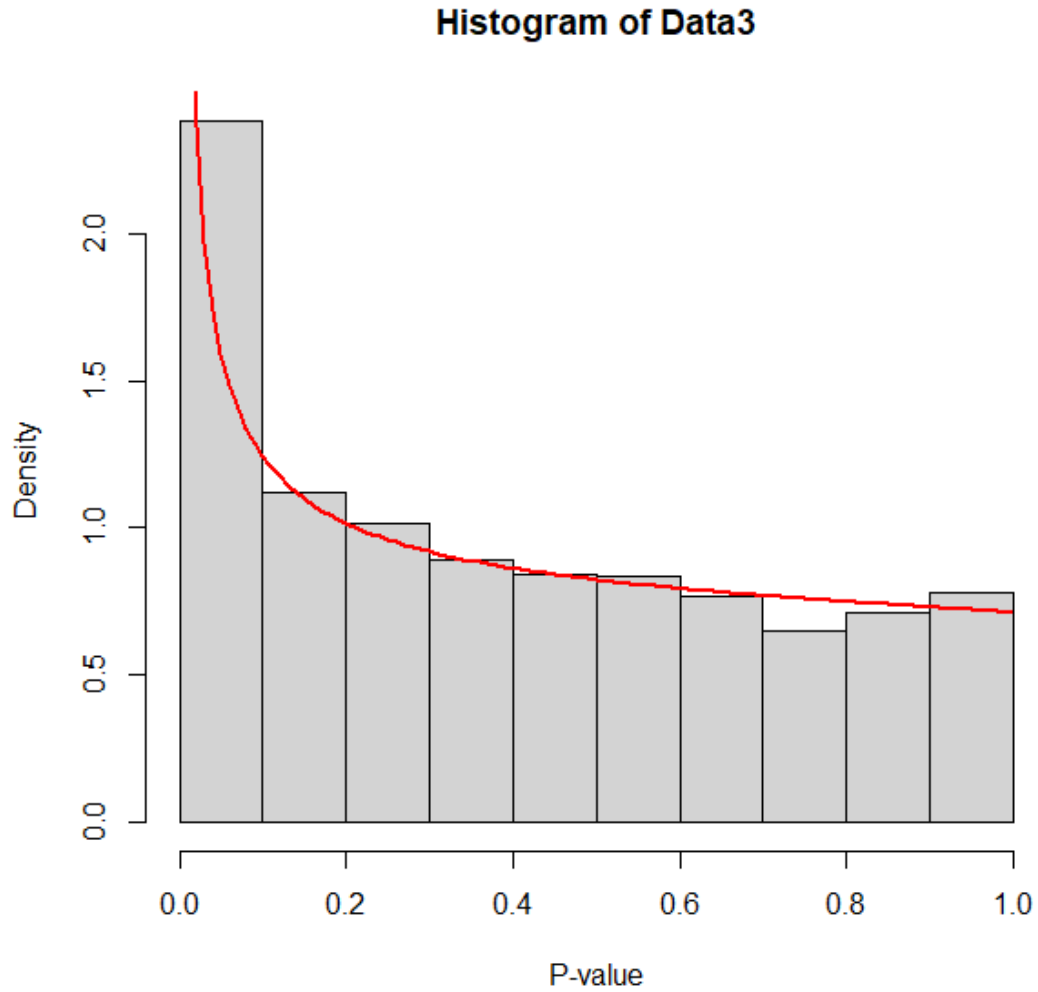


Figure 2.14: Histogram of p-values of Data 3

2.5.2 Results

Data 1:

We fitted a two-component Beta contamination model to the data 1, the fitted model is

$$0.697Beta(1, 1) + 0.303Beta(0.363, 1.997) \quad (2.11)$$

I show the fitted model with a red line on the histogram shown in Figure 2.12. And using the MLRT with constraints, we got P-value < 0.001, and the null distribution

is rejected. And according to the fitted two-component Beta contamination model, and we have 461,258 methylations in the data, the estimate $\hat{\pi} = 0.303$ indicated that about 139,761 genes were differentially expressed in the control group and Down syndrome group, and about 321497 genes were not differentially expressed.

Data 2:

We also fitted a two-component Beta contamination model to the CHR21 data (Data 2), the fitted model is

$$0.425Beta(1, 1) + 0.575Beta(0.301, 3.135) \quad (2.12)$$

I also show the fitted model with a red line on the histogram shown in Figure 2.13. And using the MLRT with constraints, we got P-value < 0.001 , and the null distribution is rejected. And according to the fitted two-component Beta contamination model, and we have 4205 methylations in the data, the estimate $\hat{\pi} = 0.575$ indicated that the gene expression levels are different in about 2418 out of 4205 genes on Chromosome 21 between control group and Down syndrome group.

Data 3:

The we fitted a two-component Beta contamination model to the data 3, the fitted model is

$$0.705Beta(1, 1) + 0.295Beta(0.339, 1.834) \quad (2.13)$$

A red line on the histogram shown shows the fitted model in Figure 2.13. And using the MLRT with constraints, we got P-value < 0.001 , and the null distribution is rejected. And according to the fitted two-component Beta contamination model, and we have 452477 methylations in the data, the estimate $\hat{\pi} = 0.295$ indicated that for data 3, the gene expression levels are different in about 133480 out of 452477 genes

between control group and Down syndrome group.

The $\hat{\pi} = 0.575$ of data 2 is larger than $\hat{\pi} = 0.303$ in the data1. It is reasonable because the proportion of differentially expressed genes on Chromosome 21 is higher on the other Chromosomes. The $\hat{\pi} = 0.295$ in data3 is the smallest because we eliminated the Down syndrome-associated genes.

Data 4:

For data 4, first we calculate the critical value based on the equation (2.10), as the sample size of all data sets $n=100$, we got critical value: $-0.00768 \times 100 + 5.1173 = 4.3493$. Fit two-component Beta contamination model with or without the constraints to the 200 data sets, and Figure 2.15 and Table 2.6 show first six examples of the histogram and fitted model. The red line is fitted Beta contamination model without constraints; the green line shows the constrained Beta contamination model, in the first example, the MLRT without constraints fail to reject the null, and the MLRT with constraints rejects the null, the example 2 to example 6, both MLRTs reject the null.

Overall, using MLRT without constraints, 164 of 200 data sets rejected the null hypothesis, and when we used MLRT with constraints adjusted for the small size, 172 rejected the null hypothesis.

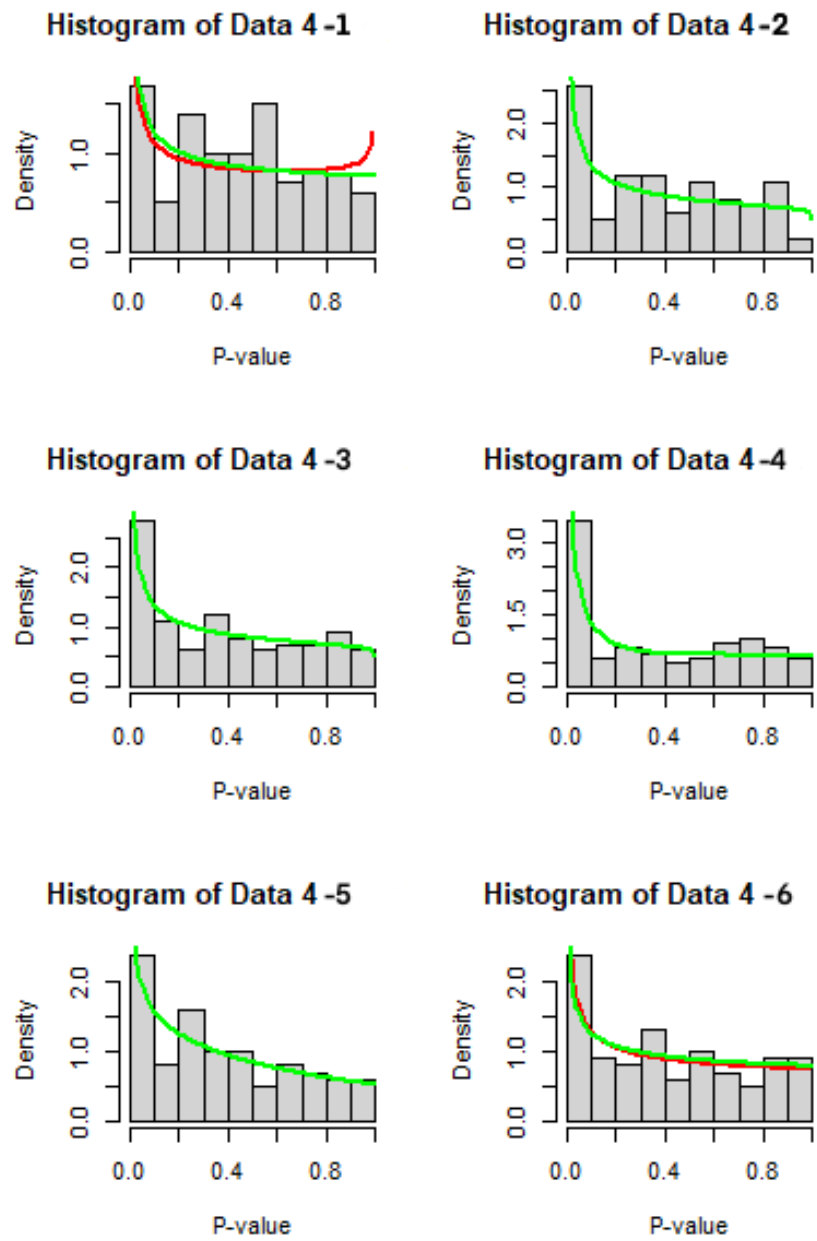


Figure 2.15: Histogram of p-values of Data 4 (6 example)

Table 2.6: Examples of fitted Beta contamination model

parameter	without constraints			with constraints		
	$\hat{\pi}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\pi}$	$\hat{\alpha}$	$\hat{\beta}$
1	0.386	0.329	0.715	0.377	0.378	1
2	0.492	0.423	1.183	0.492	0.423	1.183
3	0.480	0.426	1.215	0.480	0.426	1.215
4	0.344	0.387	7.003	0.344	0.387	7.003
5	0.479	0.665	2.096	0.479	0.665	2.096
6	0.498	0.474	0.990	0.499	0.575	1

Chapter 3 Three-component Beta Mixture Model without constraints

3.1 Introduction

A three-component Beta contamination model might better fit real data, especially when we are interested in distinguishing high differentially expressed genes and moderate differentially expressed genes in microarray data. Then testing the hypothesis of two components versus three components with a finite mixture model may be necessary. For example, assume we fit a three-component Beta contamination model to a microarray data, the fitted model is 3.1:

$$0.5Beta(1, 1) + 0.3Beta(0.7, 2) + 0.2Beta(0.2, 6) \quad (3.1)$$

The fraction of $0.5Beta(1,1)$ represents that 50% of genes are not differentially expressed, 30% of genes are moderately differentially expressed, and 20% of genes are highly differentially expressed. Suppose we could find a procedure to test two components versus three components with a Beta contamination mixture model. The microarray data could distinguish the moderate and high differential expression gene groups.

As mentioned in Chapter 1, a three-component beta mixture model has an identifiable problem. Then we considered a Beta contamination model with a kernel distribution from one parameter family by fixing the other shape parameter across all the components. When we try to find a test procedure, the LRT is a natural choice because the likelihood-based method plays a critical role in testing parametric problems and is easy to interpret. [Chen et al., 2004] introduced the modified log-likelihood ratio test(MLRT) seems to be a good choice because it has some good

asymptotic properties and is easy to apply.

In [Chen et al., 2004], they derive an MLRT to test a problem of $g=2$ versus $g \geq 3$ with a kernel distribution from a general one-parameter family. They also obtain the asymptotic null distribution of the MLRT, and it follows a mixture of χ^2 distribution. The test is relatively simple and easily applied to data with the limiting distribution. The modified log-likelihood function could be written as

$$pl(\pi, \theta_1, \theta_2, \dots, \theta_g) = \sum_{i=1}^n \log[\pi_1 f(X_i; \theta_1) + \pi_2 f(X_i; \theta_2) + \dots + \pi_g f(X_i; \theta_g)] + C_g \sum_{j=1}^g \log(\pi_j) \quad (3.2)$$

Where $C_g \sum_{j=1}^g \log(\pi_j)$ is the penalty term, $\sum_{j=1}^g \pi_j = 1$, C_g is a constant determines the penalty on the π_j . The null limiting distribution of the MLRT is not depend on the constant. The notation is in [Chen et al., 2004].

The Corollary in Chen's paper says:

Corollary 3.1.0.1. *Suppose regularity conditions 1-5 hold and that the true distribution is $f(x, G_0)$. The asymptotic distribution of the modified likelihood ratio test statistic R_n is that of the mixture*

$$\left(\frac{1}{2} - \frac{\alpha}{2\pi}\right)\chi_0^2 + \frac{1}{2}\chi_1^2 + \frac{\alpha}{2\pi}\chi_2^2 \quad (3.3)$$

where $\alpha = \cos^{-1}(\rho)$, and ρ is the correlation coefficient between the two elements of \tilde{b}_2 , which could be estimated via MMLEs.

The $f(x, G_0)$ and \tilde{b}_2 are defined in section 3.2.1 in [Chen et al., 2004].

3.2 Hypothesis Test

In [Chen et al., 2004], they considered a one parameter family. For the contaminated model, the parameters of the first component are known, so when we apply 3.1.0.1, the asymptotic distribution of the MLRT might be only an approximation.

As Beta distribution has two shape parameter, we may need to consider the following two hypotheses:

Test1: fix $\beta=1$

H_0 : p-values $\sim (1 - \pi_0)Beta(1, 1) + \pi_0Beta(\alpha_0, 1)$ versus

H_1 : p-values $\sim (1 - \pi_1 - \pi_2)Beta(1, 1) + \pi_1Beta(\alpha_1, 1) + \pi_2Beta(\alpha_2, 1)$

Test2: fix $\alpha=1$

H_0 : p-values $\sim (1 - \pi_0)Beta(1, 1) + \pi_0Beta(1, \beta_0)$ versus

H_1 : p-values $\sim (1 - \pi_1 - \pi_2)Beta(1, 1) + \pi_1Beta(1, \beta_1) + \pi_2Beta(1, \beta_2)$

We consider Test 1 first, the modified log-likelihood function is

$$l_n(\pi, \alpha_0) = \sum_{i=1}^n \log[(1 - \pi)f(X_i; 1, 1) + \pi f(X_i; \alpha_0, 1)] + C \log \pi (1 - \pi), \quad (3.4)$$

and

$$\begin{aligned} & l_n(\pi_1, \pi_2, \alpha_1, \alpha_2) \\ &= \sum_{i=1}^n \log[(1 - \pi_1 - \pi_2)f(X_i; 1, 1) + \pi_1 f(X_i; \alpha_1, 1) + \pi_2 f(X_i; \alpha_2, 1)] \\ &+ C \log(1 - \pi_1 - \pi_2) \pi_1 \pi_2, \end{aligned} \quad (3.5)$$

Use EM algorithm to get the estimate of parameters, then the MLRT can be expressed as

$$R_n = 2l_n(\hat{\pi}_1, \hat{\pi}_2, \hat{\alpha}_1, \hat{\alpha}_2) - 2l_n(\hat{\pi}, \hat{\alpha}_0) \quad (3.6)$$

Similarly, for test 2,

$$l_n(\pi, \beta_0) = \sum_{i=1}^n \log[(1 - \pi)f(X_i; 1, 1) + \pi f(X_i; 1, \beta_0)] + C \log \pi(1 - \pi), \quad (3.7)$$

and

$$\begin{aligned} & l_n(\pi_1, \pi_2, \beta_1, \beta_2) \\ &= \sum_{i=1}^n \log[(1 - \pi_1 - \pi_2)f(X_i; 1, 1) + \pi_1 f(X_i; 1, \beta_1) + \pi_2 f(X_i; 1, \beta_2)] \\ &+ C \log(1 - \pi_1 - \pi_2)\pi_1\pi_2, \end{aligned} \quad (3.8)$$

the MLRT is

$$R_n = 2l_n(\hat{\pi}_1, \hat{\pi}_2, \hat{\beta}_1, \hat{\beta}_2) - 2l_n(\hat{\pi}, \hat{\beta}_0) \quad (3.9)$$

3.3 Proof of conditions

According to the corollary 3.1.0.1, the regularity conditions need to be held when we use the asymptotic properties of MLRT. Chen, Chen, and Kalbfleisch proved that regularity conditions are held with the normal, binomial, and Poisson distribution.

If we can prove the conditions for the above beta distribution, we could apply the

modified likelihood ratio test procedure in Chen's paper. Note, we define $Y_i(\theta) =$

$$\frac{f(X_i, \theta)}{f(X_i, G_0)}, Y_i'(\theta) = \frac{f'(X_i, \theta)}{f(X_i, G_0)}, Y_i''(\theta) = \frac{f''(X_i, \theta)}{f(X_i, G_0)} \text{ and } Y_i'''(\theta) = \frac{f'''(X_i, \theta)}{f(X_i, G_0)}.$$

$f(X_i, \theta)$ is kernel density and $f(X_i, G_0)$ is probability density function of null distribution.

The following are the regularity conditions from Appendix A of [Chen et al., 2004].

Condition 1: Wald's integrability condition.

The kernel function $f(x, \theta)$ is such that the mixture distribution $f(x, G)$ satisfies

Wald's integrability conditions for consistency of the maximum likelihood estimate

(see Leroux (1992)). For this, it is sufficient to assume that

$$E|\log f(X; G_0)| < \infty$$

Condition 2: smoothness

The support of $f(x, \theta)$ is independent of θ and $f(x, \theta)$ is three times differentiable with respect to θ in Θ . Further, $f(x, \theta)$ and its derivatives with respect to θ , $f'(x, \theta)$, $f''(x, \theta)$ and $f'''(x, \theta)$ are jointly continuous in x and θ .

Condition 3: strong identifiability

For any $\theta_1 \neq \theta_2$ in Θ ,

$$\sum_{j=1}^2 a_j f(x, \theta_j) + b_j f'(x, \theta_j) + c_j f''(x, \theta_j) = 0, \text{ for all } x,$$

implies that $a_j = b_j = c_j = 0, j = 1, 2$.

Condition 4: uniform boundedness

There is an integrable function g and some $\delta > 0$ such that $|Y_i(\theta)|^{4+\delta} \leq g(X_i)$, $|Y_i'(\theta)|^3 \leq g(X_i)$, $|Y_i''(\theta)|^3 \leq g(X_i)$ and $|Y_i'''(\theta)|^3 \leq g(X_i)$ for all θ .

Condition 5: tightness

For $j = 1, 2$, the processes

$$\begin{aligned} & n^{-1/2} \sum_{i=1}^n Y_{ij}(\theta) \\ & n^{-1/2} \sum_{i=1}^n Y_i'(\theta) \\ & n^{-1/2} \sum_{i=1}^n Y_i''(\theta) \\ & n^{-1/2} \sum_{i=1}^n Y_i'''(\theta) \end{aligned} \tag{3.10}$$

are tight.

Proof:

Condition 1

As we have

$$\log f(x; G_0) = \log(1 - \pi + \pi\theta x^{\theta-1}) \leq \pi(\theta x^{\theta-1} - 1) \leq \theta x^{\theta-1}$$

and

$$\log f(x; G_0) \geq \log(1 - \pi)$$

if $\log f(x; G_0) < 0$

then $|\log f(x; G_0)| \leq -\log(1 - \pi)$

$$E|\log f(X; G_0)| \leq E[-\log(1 - \pi)] < \infty$$

if $\log f(X; G_0) \geq 0$

then $|\log f(X; G_0)| \leq \theta X^{\theta-1}$

$$E|\log f(X; G_0)| \leq E[\theta X^{\theta-1}] = \int_0^1 \theta x^{\theta-1} f(x; G_0) dx < \infty$$

Condition 2

$$f(x) = \theta x^{\theta-1}$$

$$f'(x) = x^{\theta-1}(\theta \log(x) + 1)$$

$$f''(x) = x^{\theta-1} \log(x)(\theta \log(x) + 2)$$

$$f'''(x) = x^{\theta-1} \log^2(x)(\theta \log(x) + 3)$$

are jointly continuous in x and θ .

Condition 3

$$\begin{aligned} & af(x, \theta) + bf'(x, \theta) + cf''(x, \theta) \\ &= a\theta x^{\theta-1} + bx^{\theta-1}(\theta \log(x) + 1) + cx^{\theta-1} \log(x)(\theta \log(x) + 2) \\ &= x^{\theta-1}(c\theta \log^2(x) + (2c + b\theta)\log(x) + a\theta + b) \\ &a_1 f(x, \theta_1) + b_1 f'(x, \theta_1) + c_1 f''(x, \theta_1) + a_2 f(x, \theta_2) + b_2 f'(x, \theta_2) + c_2 f''(x, \theta_2) = 0 \end{aligned}$$

implies

$$x^{\theta_1-1}[c\theta_1 \log^2(x) + (2c_1 + b_1\theta_1)\log(x) + a_1\theta_1 + b_1 + c_2\theta_2 x^{\theta_2-\theta_1} \log^2(x) + (2c_2 + b_2\theta_2)x^{\theta_2-\theta_1} \log(x) +$$

$$(a_2\theta_2 + b_2)x^{\theta_2-\theta_1}] = 0$$

as $x \in (0,1)$, $\theta_1 \neq \theta_2$, $x^{\theta-1} \neq 0$, $\log(x) \neq 0$ and $x^{\theta_2-\theta_1} \neq 0$,

if the equation above equal to 0 for all x , obviously,

$c_1\theta_1 = 0$, $2c_1 + b_1\theta_1 = 0$ and $a_1\theta_1 + b_1 = 0$, $c_2\theta_2 = 0$, $2c_2 + b_2\theta_2 = 0$ and $a_2\theta_2 + b_2 = 0$ implies $a_1 = b_1 = c_1 = a_2 = b_2 = c_2 = 0$

Condition 4

The uniform boundedness condition is satisfied for all distributions belongs to exponential family, proof see Chen's paper Appendix A. As the exponential family include beta distribution, the condition 4 is hold for the kernel we studied.

Condition 5

Based on Condition 4, consider

$$\begin{aligned} E[n^{-1/2} \sum_{i=1}^n Y_{ij}(\theta_1) - n^{-1/2} \sum_{i=1}^n Y_{ij}(\theta_2)]^2 &= E[Y_{1j}(\theta_1) - Y_{1j}(\theta_2)]^2 \\ &\leq Eg^{2/3}(X_1)|\theta_1 - \theta_2|^2 \end{aligned} \quad (3.11)$$

Then by theorem 12.3 of [Billingsley, 1968], $n^{-1/2} \sum_{i=1}^n Y_{ij}(\theta)$ is tight.[Chen et al., 2004]

similarly, we can prove $n^{-1/2} \sum_{i=1}^n Y_i'(\theta)$, $n^{-1/2} \sum_{i=1}^n Y_i''(\theta)$, and $n^{-1/2} \sum_{i=1}^n Y_i'''(\theta)$ are also tightness.

Thus, the regularity conditions are not violated with beta kernel.

■

3.4 Simulation study

We conduct extensive simulation studies on the finite beta contamination model below, and all conditions are satisfied with the kernel functions. For fixing $\alpha=1$ or $\beta=1$,

4 null and 6 alternative distributions were chosen to cover a variety of situations. The scenarios are shown below.

3.4.1 Null distributions

$\beta=1$

$$\begin{aligned}
 (N11) & 0.3Beta(1, 1) + 0.7Beta(1.5, 1) \\
 (N12) & 0.5Beta(1, 1) + 0.5Beta(0.5, 1) \\
 (N13) & 0.7Beta(1, 1) + 0.3Beta(0.5, 1) \\
 (N14) & 0.7Beta(1, 1) + 0.3Beta(3, 1)
 \end{aligned} \tag{3.12}$$

$\alpha = 1$

$$\begin{aligned}
 (N21) & 0.3Beta(1, 1) + 0.7Beta(1, 1.5) \\
 (N22) & 0.5Beta(1, 1) + 0.5Beta(1, 0.5) \\
 (N23) & 0.7Beta(1, 1) + 0.3Beta(1, 0.5) \\
 (N24) & 0.7Beta(1, 1) + 0.3Beta(1, 3)
 \end{aligned} \tag{3.13}$$

3.4.2 Alternative distributions

$\beta=1$

$$\begin{aligned}
 (A11) & 0.3Beta(1, 1) + 0.35Beta(1.5, 1) + 0.35Beta(0.5, 1) \\
 (A12) & 0.6Beta(1, 1) + 0.2Beta(1.5, 1) + 0.2Beta(0.5, 1) \\
 (A13) & 0.6Beta(1, 1) + 0.2Beta(0.2, 1) + 0.2Beta(0.8, 1) \\
 (A14) & 0.6Beta(1, 1) + 0.2Beta(2, 1) + 0.2Beta(3, 1) \\
 (A15) & 0.6Beta(1, 1) + 0.2Beta(1.5, 1) + 0.2Beta(8, 1) \\
 (A16) & 0.8Beta(1, 1) + 0.1Beta(0.2, 1) + 0.1Beta(0.8, 1)
 \end{aligned} \tag{3.14}$$

$\alpha=1$

$$\begin{aligned}(A21) & 0.3Beta(1, 1) + 0.35Beta(1, 1.5) + 0.35Beta(1, 0.5) \\(A22) & 0.6Beta(1, 1) + 0.2Beta(1, 1.5) + 0.2Beta(1, 0.5) \\(A23) & 0.6Beta(1, 1) + 0.2Beta(1, 0.2) + 0.2Beta(1, 0.8) \\(A24) & 0.6Beta(1, 1) + 0.2Beta(1, 2) + 0.2Beta(1, 3) \\(A25) & 0.6Beta(1, 1) + 0.2Beta(1, 1.5) + 0.2Beta(1, 8) \\(A26) & 0.8Beta(1, 1) + 0.1Beta(1, 0.2) + 0.1Beta(1, 0.8)\end{aligned}\tag{3.15}$$

We performed 5000 repetitions in the simulations and considered the significance level 0.05. Since the data sets are simulated, we use the true value as the initial point when fitting the mixture model to obtain the maximum of the log-likelihood function, and all the initial weights are set to be equal. That makes the convergence time of the EM algorithm to be shorter. After investigation, we found that if we generate data from different scenarios and use EM algorithm, in 78% cases, the difference between the true value and the estimated maximum log-likelihood are smaller than 5%; in 22% cases, the difference are larger than 5% of the true value, but the value still not far from the true value. It is better to set different initial values to increase the probability of obtaining the global optima in applications. If we set 5 sets of initial values, 96% of all the data have difference less than 5% between true and estimated maximum log-likelihood.

We also use the bootstrap method as a competing method to compare using the asymptotic properties of null limiting distribution. The computing algorithm is as follows. Part of the notation and criteria considered [Chen et al., 2004] section 4.2.1 as reference.

Step 1: Draw a sample with size n from the null distribution and obtaining the

MMLEs of the two-component beta contamination model and three-component beta contamination mode via the EM algorithm.

Step 2: Computing the MLRT statistic R_n .

Step 3: Draw m bootstrap sample of size n from the two-component contamination model with parameter we obtained from MMLEs in previous step. Calculating the MMLEs for the bootstrap sample under the null and alternative model hypothesis respectively.

Step 4: Calculating the MLRT statistic R_n^* for each bootstrap sample.

Step 5: Calculating the $100(1 - \alpha)\%$ percentile of the MLRT statistics $R_n^{*(1)}, R_n^{*(2)}, \dots, R_n^{*(m)}$ of each bootstrap sample, then compare it with MLRT statistic R_n we obtained from the original null distribution. If statistic R_n is larger, reject the null hypothesis.

The bootstrap method is much different than the asymptotic test: When the components increase to three, the time of EM algorithm converge becomes much longer than a two-component mixture model, with hundreds of repetitions, the time cost would be quite long, and we also set more initial points to increase the probability of obtaining the global maximum. Thus, we set our bootstrap size at 1000 and use 5 sets of initial values.

3.4.3 Actual rejection rates and powers

Figure 3.1 and 3.2 show the actual rejection rate when $\beta=1$ or $\alpha=1$ and Figure 3.3 and 3.4 show the power curves.

When we look at the simulation results, we found when we fix α or β , the simulation results are almost the same except for some random error produced by computation. That was because the properties of beta distribution with one parameter fixed to 1. For example the $0.5\text{Beta}(1,1)+0.5\text{Beta}(1,0.5)$ shares totally same but symmetric

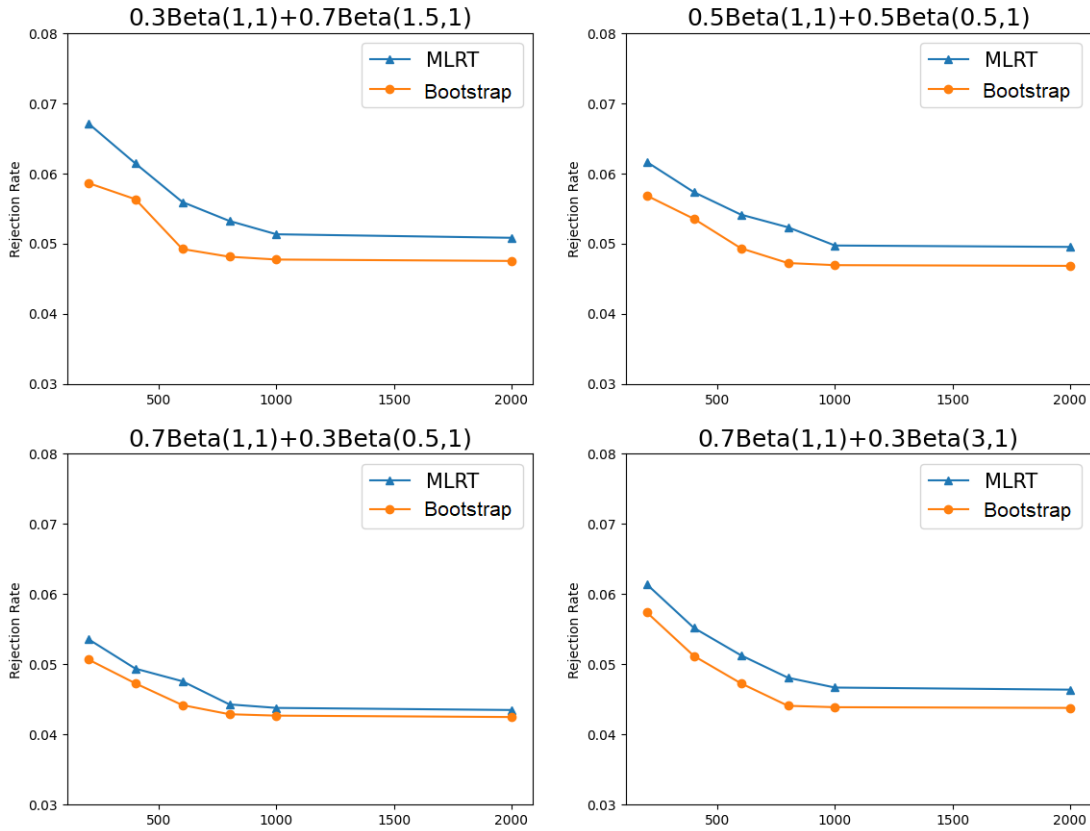


Figure 3.1: Actual rejection rate when $\beta=1$

shape with $0.5\text{Beta}(1,1)+0.5\text{Beta}(0.5,1)$, see the Figure 3.5. Therefore, we look at the results of $\beta=1$.

When we look at Figure 3.1, the actual rejection rate of all the four scenarios is not far from 0.05. As the sample size increases, the actual rejection rate decreases and becomes consistent. When the weights of contamination become smaller, the actual rejection rate goes slightly under 0.05.

As shown in Figure 3.3, the performance of the bootstrap method is better than

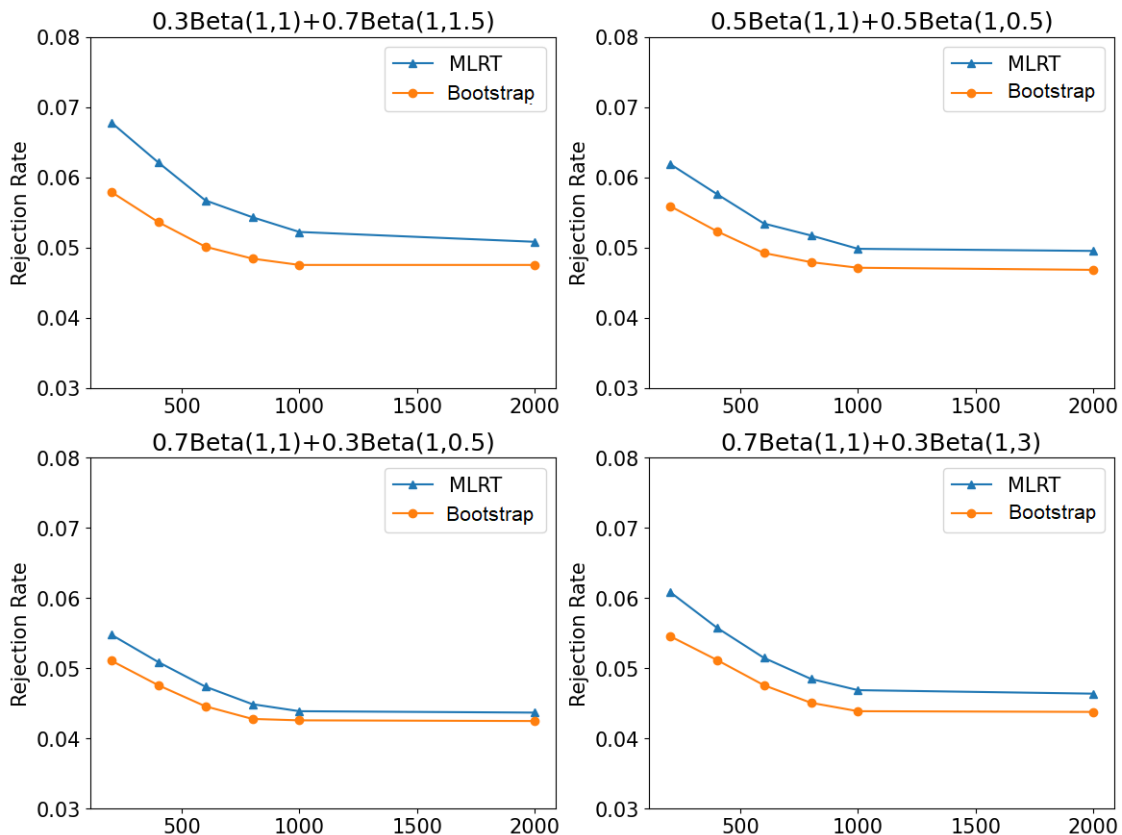


Figure 3.2: Actual rejection rate when $\alpha=1$

the MLRT in general, but the time cost of the bootstrap method is much longer compared to the MLRT, especially when the sample size is getting larger. We mentioned in section 3.2, as the parameter of first component in contaminated model are known, the asymptotic distribution in Corollary3.1.0.1 may be approximation. Based on the simulation result, the performance of the approximate asymptotic property is good.

The power of scenario (A11) is higher than (A12), and the power of (A13) is larger than (A16); the finding makes sense because as the weights of Beta(1,1) increase, it is harder to distinguish the two contamination. Scenarios (A12) and (A15) have higher

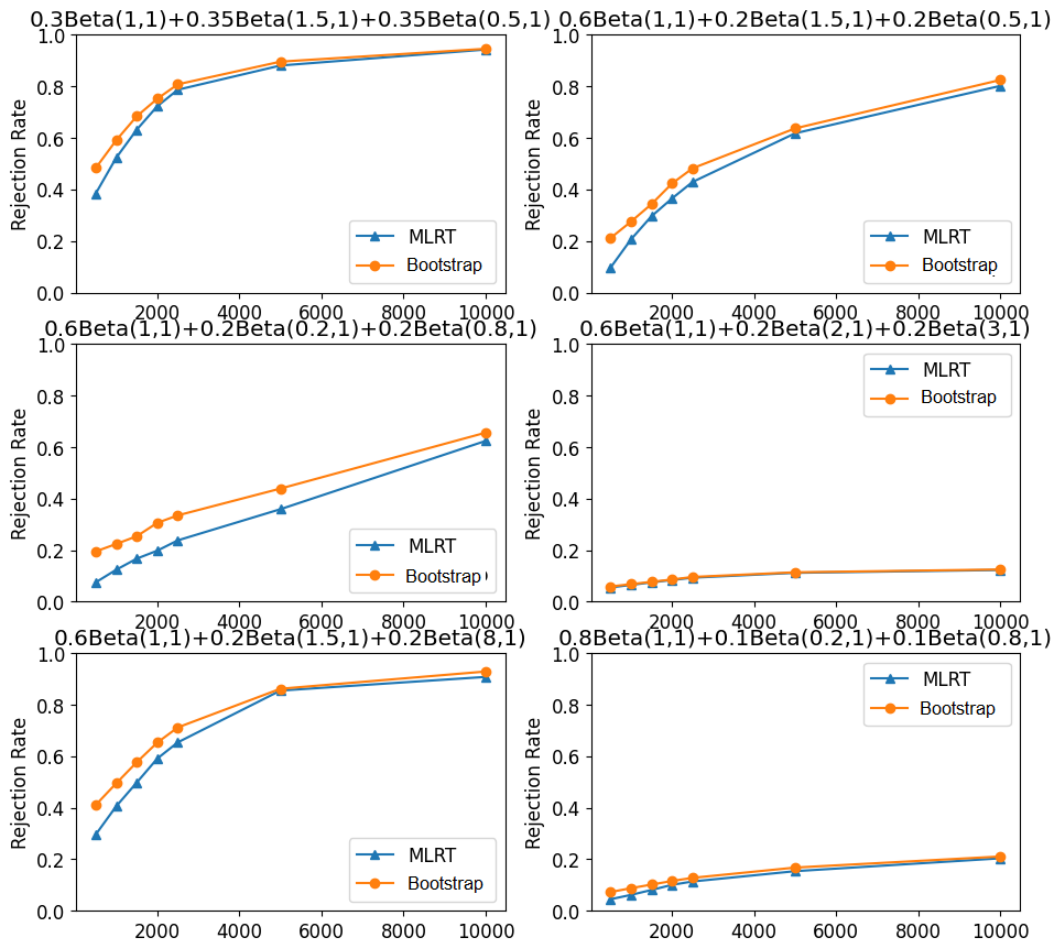


Figure 3.3: Power curves when $\beta=1$

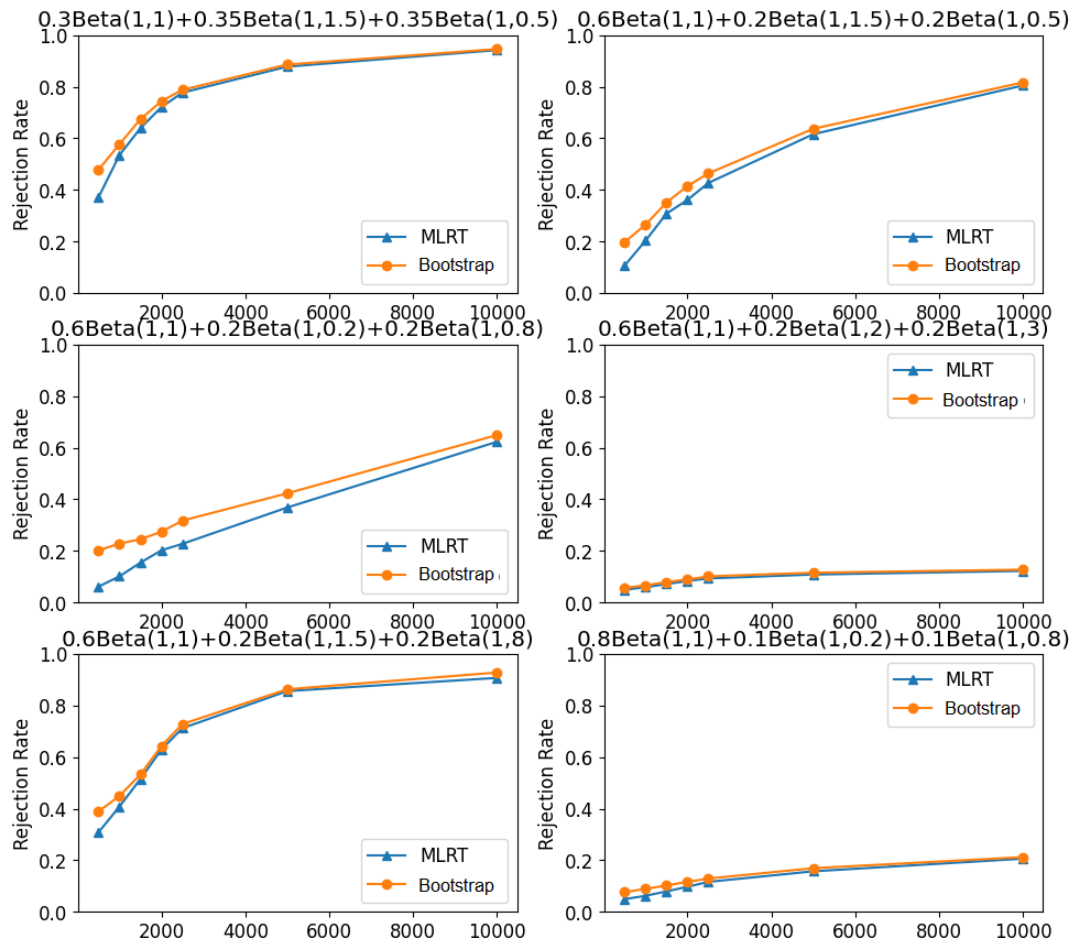


Figure 3.4: power curves when $\alpha=1$

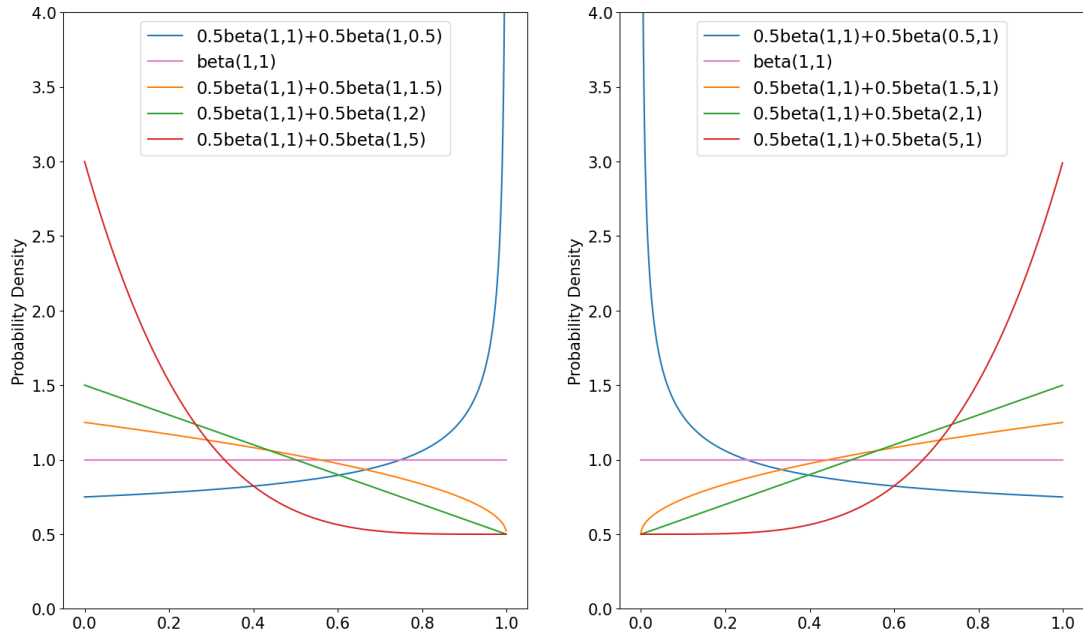


Figure 3.5: Basic Shape of two-component Beta contamination model

power than (A13). It is plausible when we look at the plots of (A12) and (A15) in figure 3.7. The tail of (A12) distribution gets thicker as it is close to 1, and the left tail of (A15) has a drop close to 0. The distribution shapes of (A12) and (A15) are too complicated to have a good fit with a two-component beta contamination model when one shape parameter is fixed. Thus, the null is easy to be rejected. Compared to (A14), when samples were generated from the mixture model with β_1 and $\beta_2 > 1$ and quite close, the differentiation in the distributions of two contamination is tiny; thus, a two-component model could give a good fit, in this case, the power turns to be quite small.

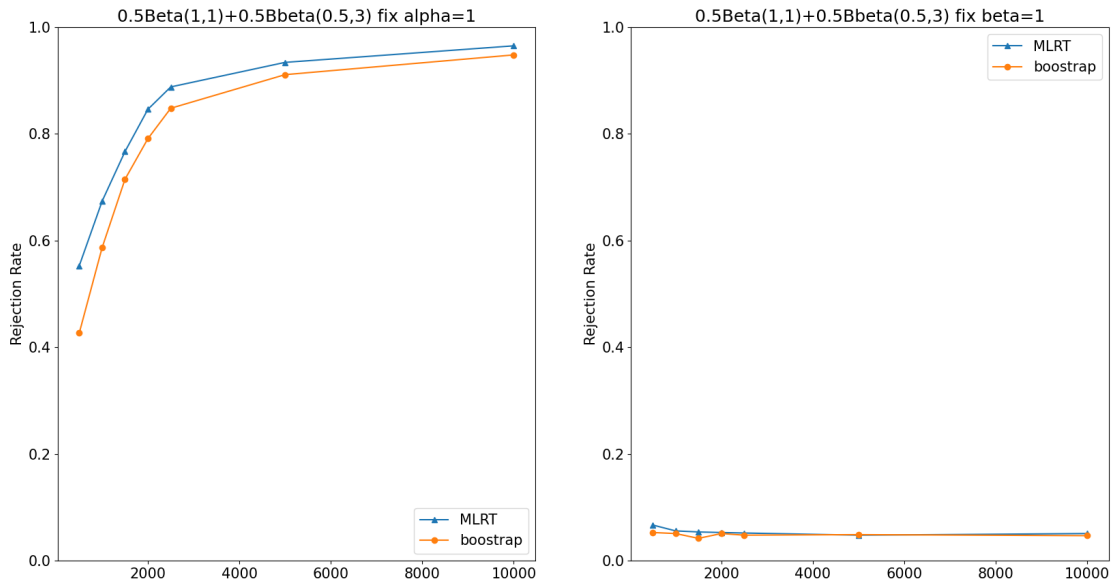


Figure 3.6: Power curves of sample generate from $0.5\text{Beta}(1,1)+0.5\text{Beta}(0.5,3)$

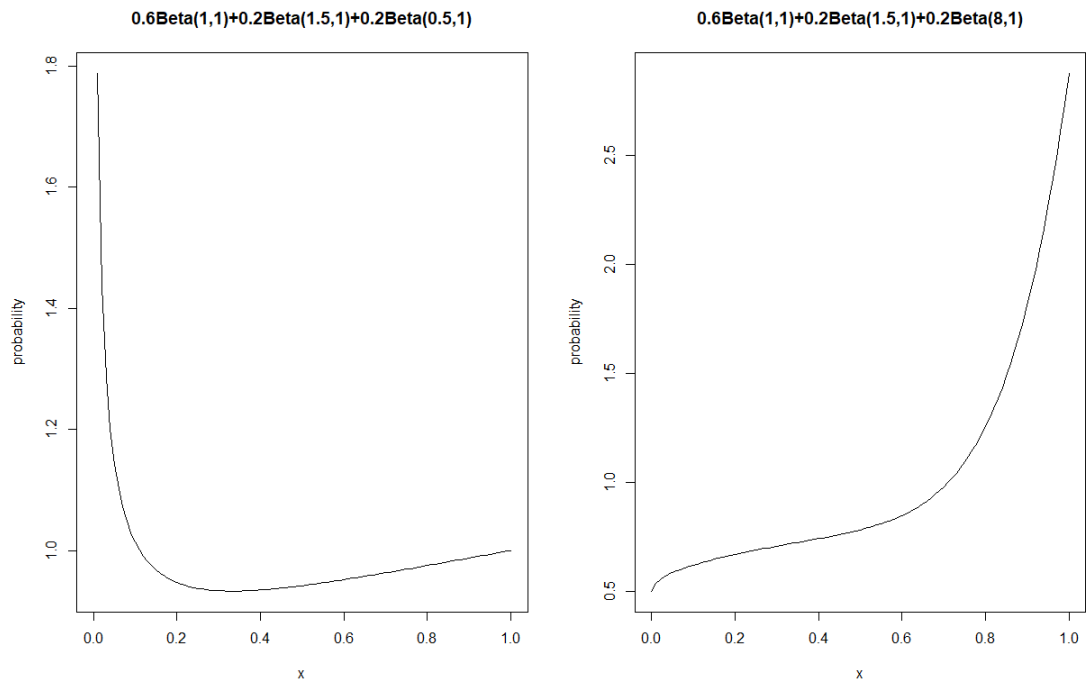


Figure 3.7: Density plot of (A12) and (A15)

Sample generate from $0.5\text{Beta}(1,1)+0.5\text{Beta}(0.5,3), \alpha=1$

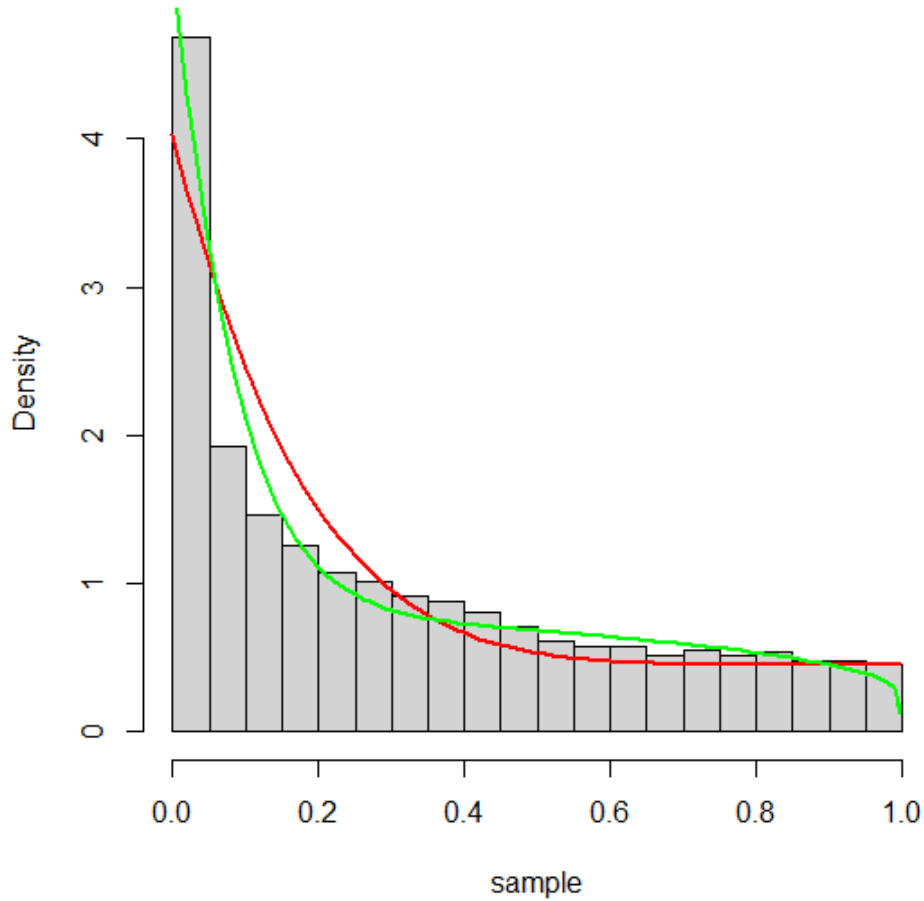


Figure 3.8: The Histogram and fitted model when fix $\alpha=1$, the red line show 2-component CB model, the green line show 3-component CB model

3.4.4 Problems in MLRT with one parameter family

The first problem is when the sample size is not large enough, the power obtained from the bootstrap method and MLRT are too low to ensure the tests are efficient.

The second problem is that the beta contamination model with one parameter family is not sufficient to describe a variety of shapes of distribution accurately. That may result in some problems.

Sample from $0.5\text{Beta}(1,1)+0.5\text{Beta}(0.5,3),\beta=1$

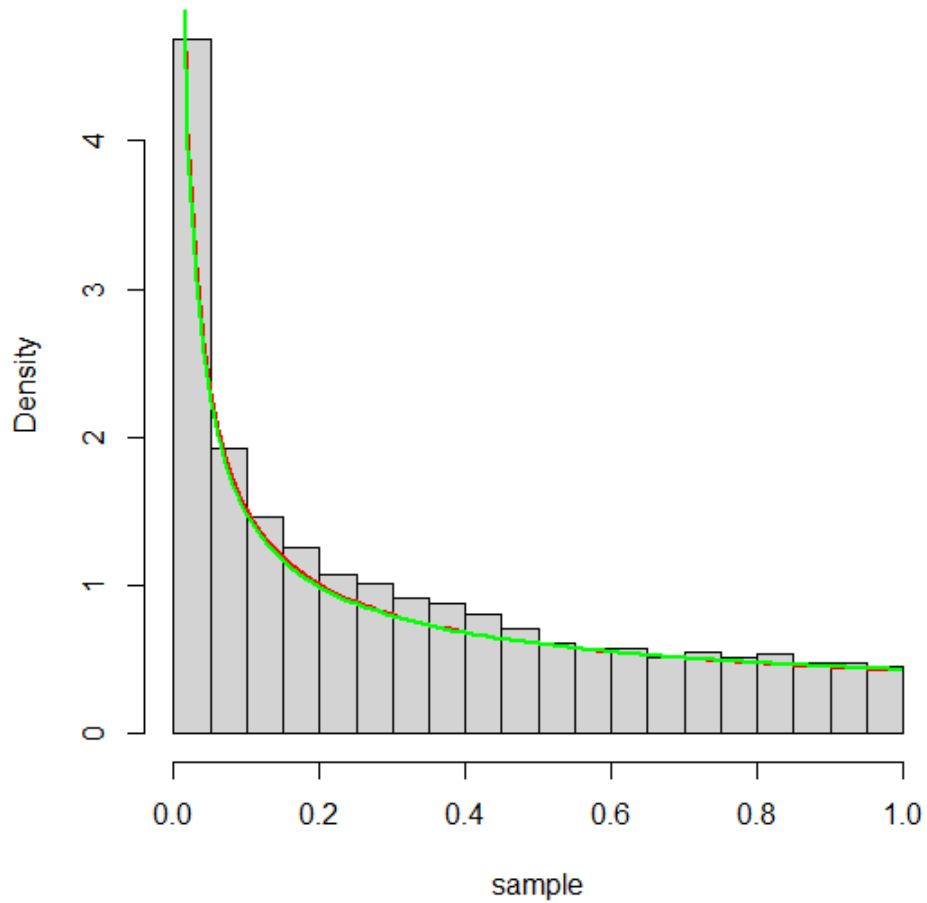


Figure 3.9: The Histogram and fitted model when fix $\beta=1$, the red line show 2-component CB model, the green line show 3-component CB model

For all the simulations results shown above, the samples were generated from a beta contamination model of a one-parameter family; in other words, we fixed one of the shape parameters to be equal to 1.

Now we generated samples from $0.5\text{Beta}(1,1)+0.5\text{Beta}(0.5,3)$, and use the MLRT and bootstrap method to compute powers as we did before, the power curves are shown in figure 3.6.

As figure 3.6 shows, when we use the test with $\alpha = 1$, although the true model has two-component, the rejection rates of tests are quite high. Figure 3.8 show the histogram of a sample generated from $0.5\text{Beta}(1,1)+0.5\text{Beta}(0.5,3)$ and plot fitted two-component model with red line, three-component model with green line. When we fix shape parameter α , a two-component beta contamination model with β parameter family could not give a precise fit to the data we generated from $0.5\text{Beta}(1,1)+0.5\text{Beta}(0.5,3)$. In other words, a three-component beta contamination model almost always gives a better fit than a two-component model since it has more freedom. With this model, we may make large type I error.

When we use the test with β fixed, the power curve are shown in figure 3.6: in most cases, the null hypothesis was not rejected. And when we look at figure 3.9, the fitted beta contamination model with α parameter family seems fit the data properly. The fitted two-component beta contamination model

$$0.101\text{Beta}(1, 1) + 0.899\text{Beta}(0.36, 1) \quad (3.16)$$

The fitted three-component beta contamination model

$$0.102\text{Beta}(1, 1) + 0.448\text{Beta}(0.28, 1) + 0.450\text{Beta}(0.45, 1) \quad (3.17)$$

But when we look at the fitted model, the problem is that the beta contamination model with the α parameter family tends to over-estimate the weights of two contamination. Applying this test to the real data will underestimate the proportion of genes that are not differentially expressed.

Therefore, as the test with one parameter family have some problem in some cases, when test $g=2$ vs. $g=3$, we may consider a beta contamination model with a two-

parameter family in the next chapter.

3.5 Real Data Application

3.5.1 Introduction

We continue to the microarray data we used in the previous chapters: data on the systematic genome-wide DNA methylation alternation in blood cells of toddlers with Down Syndrome. 34 children with age 0.5–4.5 years take part in this study, 17 has Down syndrome, and 17 are typically developing children[Naumova et al., 2021]. The data is available on the website below:

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE174555>.

Data 1, Data 2 and Data 3 are same with Chapter 2. Introductions are in section 2.5.1.

For MLRT, ρ in 3.1.0.1 could be estimated by the MMLEs of the null distribution; thus, we could get the null limiting distribution, which is a mixture of χ^2 , reject the null if

$$P\left(\left(\frac{1}{2} - \frac{\alpha}{2\pi}\right)\chi_0^2 + \frac{1}{2}\chi_1^2 + \frac{\alpha}{2\pi}\chi_2^2 > R_n\right) < 0.05 \quad (3.18)$$

For the bootstrap method, we set the bootstrap size as 1000, use the same steps used in section 3.4, and reject the null if R_n is larger than .95 quantile.

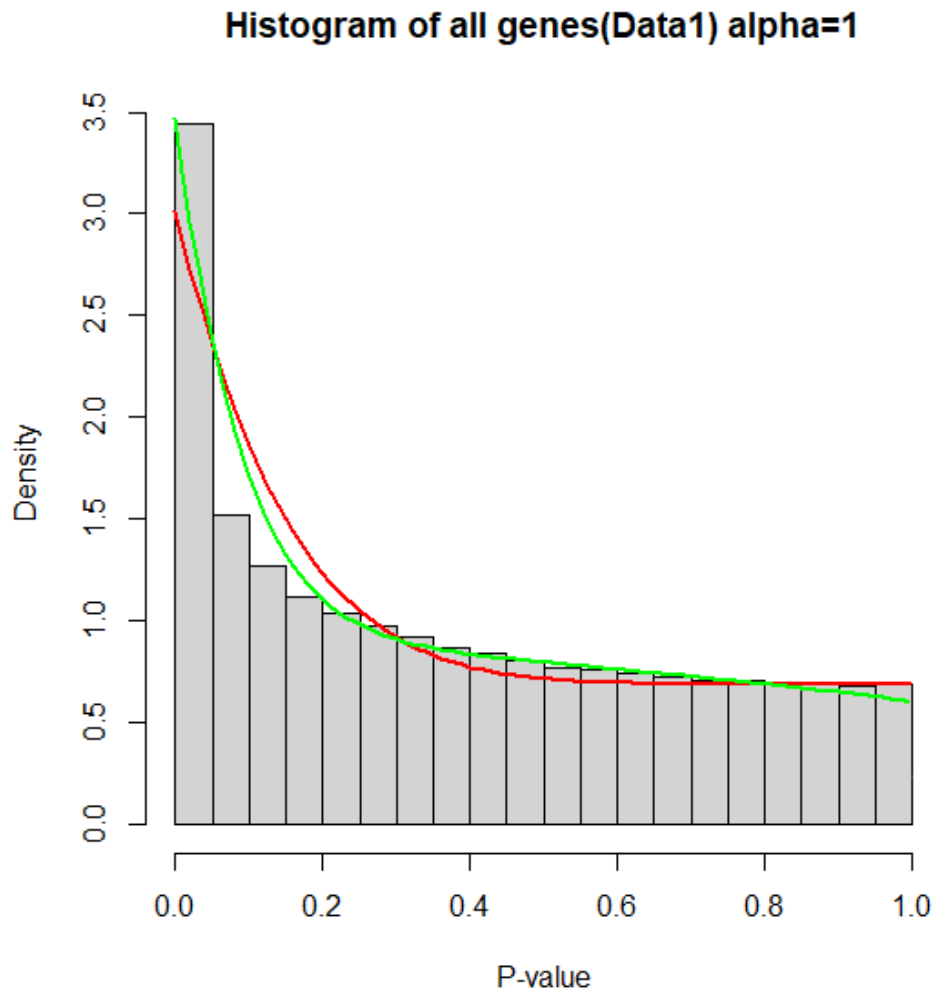


Figure 3.10: Histogram of p-values in Data 1 with $\alpha = 1$, the red line shows the two-component fitted model, and the green line shows the three-component fitted model.

Histogram of all genes(Data1) beta=1

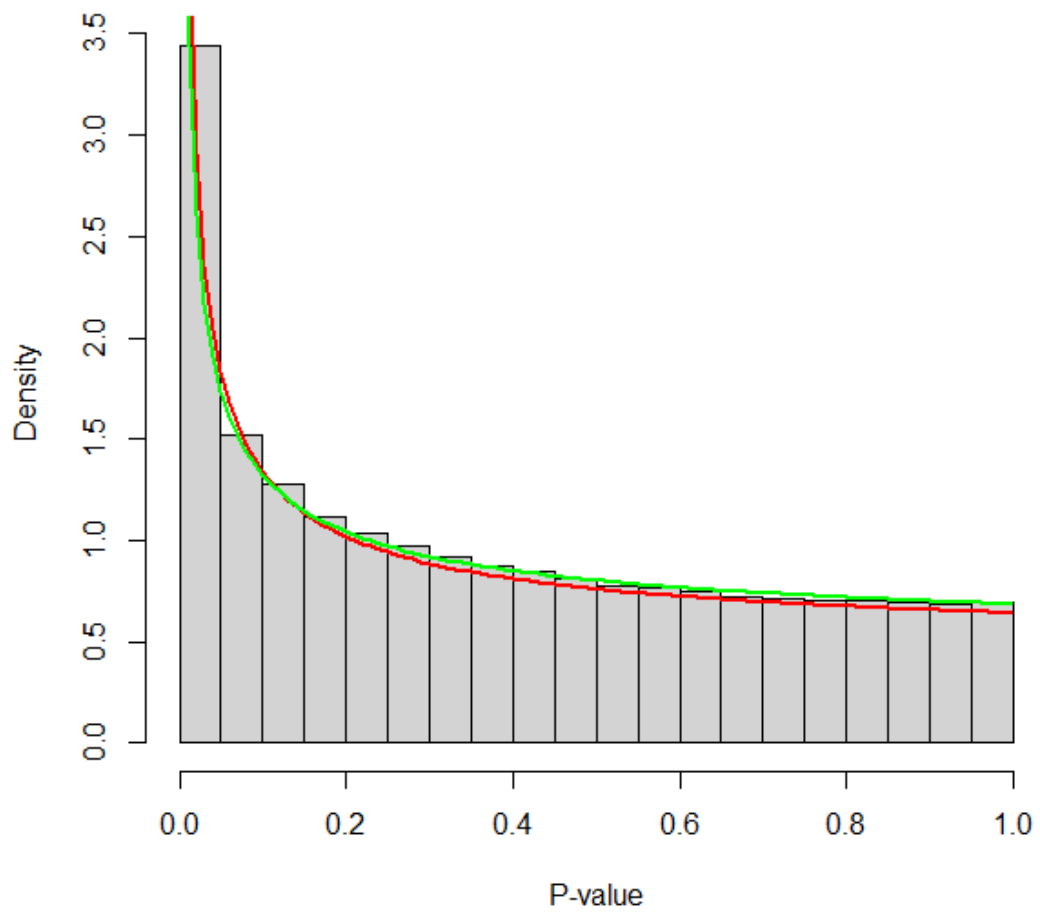


Figure 3.11: Histogram of p-values in Data 1 with $\beta = 1$, the red line shows the two-component fitted model, and the green line shows the three-component fitted model.

3.5.2 Results

Data1: We fitted a two-component Beta Contamination model to data 1 with a fixed $\alpha=1$, the fitted model is

$$0.677Beta(1, 1) + 0.363Beta(1.7, 0.45) \quad (3.19)$$

We also fit a three-component Beta contamination model to data 1, and obtain a fitted model below

$$0.603Beta(1, 1) + 0.181Beta(1, 1.851) + 0.216Beta(1, 11.734) \quad (3.20)$$

The histogram is shown in Figure 3.10, the red line shows the two-component fitted model, and the green line shows the three-component fitted model. The other problem is that, when we fix α , the proportion of Beta(1,1) for the two-component and three-component model are different in most case, it makes this the model difficult to interpret.

As discussed in the last section, the three-component model always gives a better fit, if we use this model, the type I error (reject the true H_0) might be high.

On the other hand, we fitted a two-component Beta Contamination model to data 1 with a fixed $\beta=1$; the fitted model is

$$0.408Beta(1, 1) + 0.592Beta(0.393, 1) \quad (3.21)$$

And the fitted three-component Beta contamination model to data 1 is

$$0.408Beta(1, 1) + 0.306Beta(0.350, 1) + 0.286Beta(0.613, 1) \quad (3.22)$$

The histogram is shown in Figure 3.11, the red line shows the two-component fitted model, and the green line shows the three-component fitted model.

Both methods reject the null hypothesis, and according to the fitted three-component Beta contamination model and we have 461,258 methylations in the data, the estimate $\hat{\pi}_1 = 0.306$ and $\hat{\pi}_2 = 0.286$ indicated that about 141145 genes were highly differentially expressed of the control group and Down syndrome group, about 123617 genes were moderate differentially expressed of the control group and Down syndrome group, and about 188193 genes are not differentially expressed.

But if we compare the fitted model with the two-component model we fit in chapter 2, the proportion of genes that are not differentially expressed is 40.8, which is much lower than 69.7 we obtained in chapter 2. Which indicates the model might overestimates the proportion of differentially expressed gene since the model with one parameter family lack freedom.

Data 2:

We also fitted a two-component Beta contamination model to the CHR21 data (Data 2) with $\beta=1$, the fitted model is

$$0.111Beta(1, 1) + 0.889Beta(0.331, 1) \quad (3.23)$$

The fit a three-component Beta contamination model with $\beta=1$ to data 2, the fitted model is

$$0.111Beta(1, 1) + 0.437Beta(0.213, 1) + 0.452Beta(0.514, 1) \quad (3.24)$$

The histogram is shown in Figure 3.12, the red line shows the two-component fitted

Histogram of p-values of CHR 21(Data2)

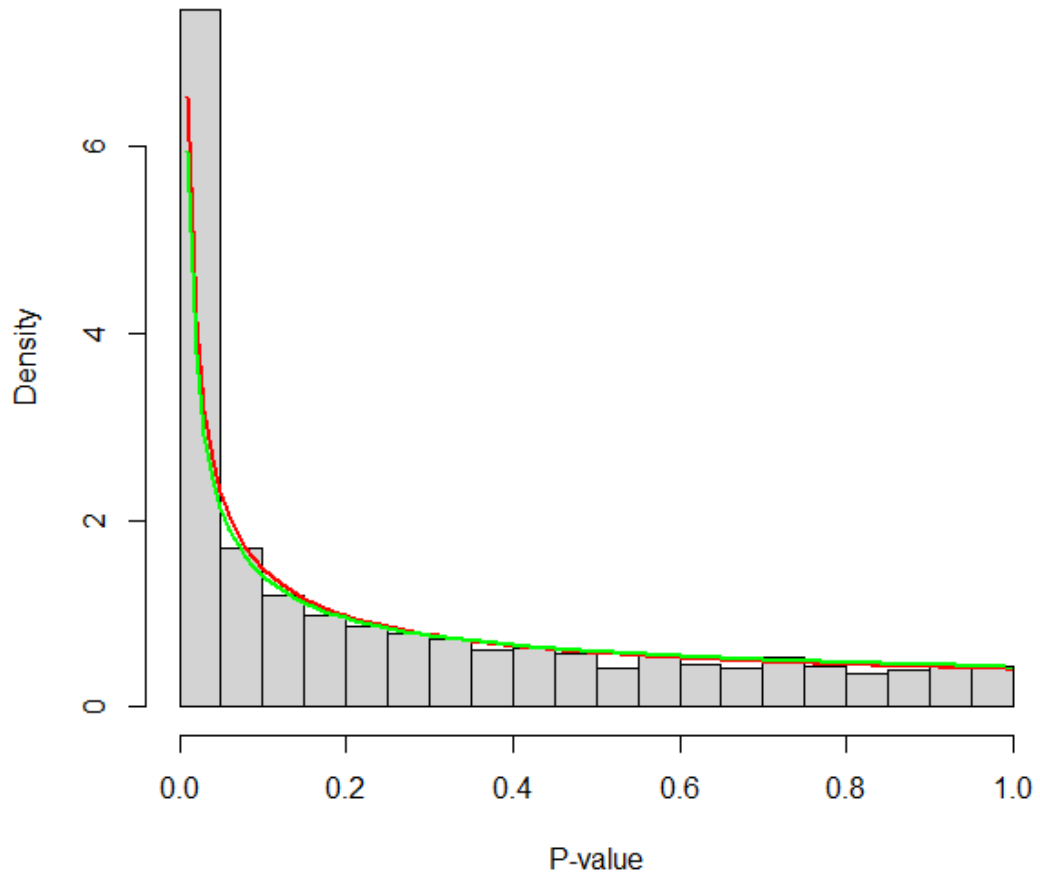


Figure 3.12: Histogram of p-values in Data 2 with $\beta = 1$, the red line shows the two-component fitted model, and the green line shows the three-component fitted model.

model, and the green line shows the three-component fitted model.

Both methods fail to reject the null, when we use 3.1.0.1, the p-value is 0.22, when we use bootstrap, the p-value is 0.19. That indicates the p values follow a two-component Beta contamination model. According to the fitted two-component Beta contamination model, and we have 4205 methylations in data 2, the estimate $\hat{\pi} = 0.889$ indicated that the gene expression levels are different in about 3738 out of 4205 genes on Chromosome 21 between control group and Down syndrome group.

Compared to the weight of 0.575 we got in chapter 2, this model has a risk of over-estimating the number of differentially expressed genes.

Data 3:

The fitted model of a two-component Beta contamination model with $\beta=1$ is

$$0.414Beta(1, 1) + 0.576Beta(0.429, 1) \quad (3.25)$$

And the fitted constrained three-component Beta contamination model is:

$$0.414Beta(1, 1) + 0.381Beta(0.392, 1) + 0.205Beta(0.737, 1) \quad (3.26)$$

A histogram shows the fitted model in Figure 3.13, the red line shows the two-component fitted model, and the green line shows the three-component fitted model.

Both methods reject the null, and according to the fitted three-component Beta contamination model, and we have 452,477 methylations in the data 3, the estimate $\hat{\pi}_1 = 0.381$ and $\hat{\pi}_2 = 0.205$ indicated that about 172394 genes were highly differentially expressed of the control group and Down syndrome group, about 92758

Histogram of Data3

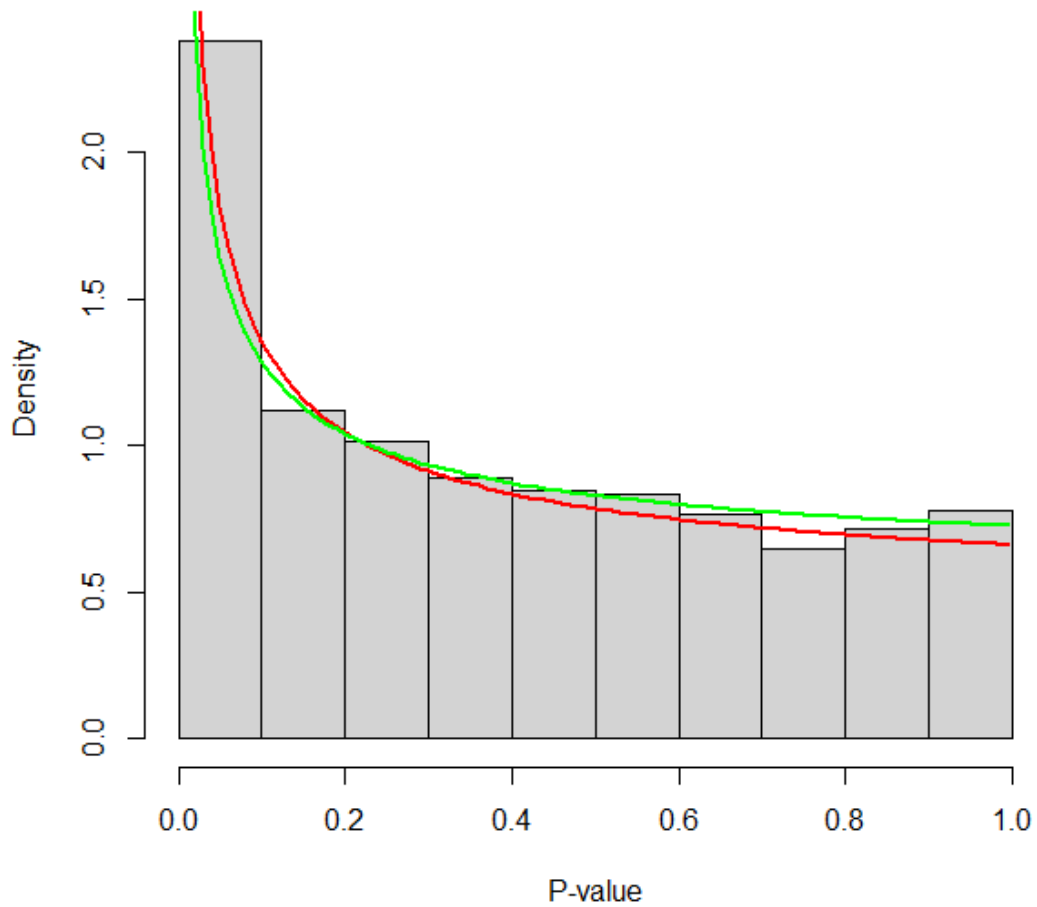


Figure 3.13: Histogram of p-values in Data 3 with $\beta = 1$, the red line shows the two-component fitted model, and the green line shows the three-component fitted model.

genes were moderate differentially expressed of the control group and Down syndrome group, and about 187325 genes are not differentially expressed.

Compared to the MMLEs in chapter 2, the estimation here deviated from the truth.

As the simulation and real data analysis have shown, the beta contamination model with one parameter family seems not to be a good choice to test $k=2$ vs. $k=3$ in the microarray data analysis; we will discuss a new method in chapter 4.

Chapter 4 Three-component Beta Mixture Model with Constraints

4.1 Introduction

Consider

$$P_1, \dots, P_n \stackrel{iid}{\sim} (1 - \gamma)Beta(1, 1) + \gamma Beta(\alpha, \beta) \quad (4.1)$$

where $\gamma \in [0, 1]$, $\alpha \in (0, \infty)$ and $\beta \in (0, \infty)$. We have mentioned in Chapter 1, assuming p-values are independent and identically distributed, the two-component Beta contamination model describes p-values in microarray data analysis, then we can apply homogeneity hypothesis tests as we did in Chapter 2.

We also state that in some cases, a two-component beta mixture model is not sufficient since the distribution of P-values is complicated or researchers are interested in the more detailed information in the microarray data. So we studied the top of using a Beta contamination model with one parameter family in Chapter 3. Due to the limitations of the Beta contamination model with one parameter family, in this chapter, we will study a Beta contamination model with a two-parameter family.

Consider a three-component beta mixture model:

$$\gamma_1 Beta(\alpha_1, \beta_1) + \gamma_2 Beta(\alpha_2, \beta_2) + \gamma_3 Beta(\alpha_3, \beta_3) \quad (4.2)$$

where $\gamma_j \in [0, 1]$ and $\sum_{j=1}^3 \gamma_j = 1$, $\alpha_j \in (0, \infty)$ and $\beta_j \in (0, \infty)$, $j=1, 2, 3$.

As we exemplified in Chapter 1 example 1.2, what appears to be a 3-component model can be expressed as a uniform distribution. So we need to find some con-

straints to guarantee the identifiability of the 3-component Beta mixture model. In other words, we need to make sure a 3-component Beta mixture model with constraints cannot be expressed as a different 3-component Beta mixture model nor reduced to a 2-component model.

4.2 Identifiability of Beta Mixture Models

4.2.1 Identification of Three-component Beta Mixture Model

Let

$$f(x) = \gamma_1 B(\alpha_1, \beta_1) x^{\alpha_1-1} (1-x)^{\beta_1-1} + \gamma_2 B(\alpha_2, \beta_2) x^{\alpha_2-1} (1-x)^{\beta_2-1} + \gamma_3 B(\alpha_3, \beta_3) x^{\alpha_3-1} (1-x)^{\beta_3-1} \quad (4.3)$$

and

$$g(x) = \delta_1 B(\theta_1, \eta_1) x^{\theta_1-1} (1-x)^{\eta_1-1} + \delta_2 B(\theta_2, \eta_2) x^{\theta_2-1} (1-x)^{\eta_2-1} + \delta_3 B(\theta_3, \eta_3) x^{\theta_3-1} (1-x)^{\eta_3-1} \quad (4.4)$$

Theorem 4.2.1. *A 3-component Beta mixture model shown in 4.3 with $\gamma_i \in (0, 1)$, $\alpha_1 < \alpha_2 < \alpha_3 \in (0, \infty)$ and $\beta_1 < \beta_2 < \beta_3 \in (0, \infty)$ cannot be expressed as a 3-component beta mixture show as 4.4 with $\delta_i \in [0, 1]$, $\theta_1 \leq \theta_2 \leq \theta_3 \in (0, \infty)$ and $\eta_1 \leq \eta_2 \leq \eta_3 \in (0, \infty)$ unless $\gamma_1 = \delta_1$, $\alpha_1 = \theta_1$, $\beta_1 = \eta_1$, $\gamma_2 = \delta_2$, $\alpha_2 = \theta_2$, $\beta_2 = \eta_2$, $\gamma_3 = \delta_3$, $\alpha_3 = \theta_3$ and $\beta_3 = \eta_3$.*

Proof:

Suppose $f(x) = g(x)$ and let $B(\alpha_i, \beta_i) = C_i$ and $B(\theta_j, \eta_j) = K_j$

Assume that $\alpha_1 < \alpha_2 < \alpha_3$, $\beta_1 < \beta_2 < \beta_3$, $\theta_1 \leq \theta_2 \leq \theta_3$ and $\eta_1 \leq \eta_2 \leq \eta_3$.

Multiply by $x^{1-\alpha_1}$ and take limits of both sides, we have

$$\lim_{x \rightarrow 0} x^{1-\alpha_1} f(x) = \lim_{x \rightarrow 0} x^{1-\alpha_1} g(x) \quad (4.5)$$

implies

$$\gamma_1 C_1 = \begin{cases} \infty & \text{if } \theta_1 \text{ or } \theta_2 \text{ or } \theta_3 < \alpha_1 \\ 0 & \text{if } \theta_1 \text{ and } \theta_2 \text{ and } \theta_3 > \alpha_1 \\ \delta_1 K_1 & \text{if } \alpha_1 = \theta_1 < \theta_2 \leq \theta_3 \\ \delta_1 K_1 + \delta_2 K_2 & \text{if } \alpha_1 = \theta_1 = \theta_2 < \theta_3 \\ \delta_1 K_1 + \delta_2 K_2 + \delta_3 K_3 & \text{if } \alpha_1 = \theta_1 = \theta_2 = \theta_3 \end{cases} \quad (4.6)$$

We get contradiction if θ_1 or θ_2 or $\theta_3 < \alpha_1$ and if θ_1 and θ_2 and $\theta_3 > \alpha_1$.

Then, multiply the equation $f(x) = g(x)$ by $(1-x)^{1-\beta_1}$ and take limits of both sides, we have

$$\lim_{x \rightarrow 1} (1-x)^{1-\beta_1} f(x) = \lim_{x \rightarrow 1} (1-x)^{1-\beta_1} g(x) \quad (4.7)$$

implies

$$\gamma_1 C_1 = \begin{cases} \infty & \text{if } \eta_1 \text{ or } \eta_2 \text{ or } \eta_3 < \beta_1 \\ 0 & \text{if } \eta_1 \text{ and } \eta_2 \text{ and } \eta_3 > \beta_1 \\ \delta_1 K_1 & \text{if } \beta_1 = \eta_1 < \eta_2 \leq \eta_3 \\ \delta_1 K_1 + \delta_2 K_2 & \text{if } \beta_1 = \eta_1 = \eta_2 < \eta_3 \\ \delta_1 K_1 + \delta_2 K_2 + \delta_3 K_3 & \text{if } \beta_1 = \eta_1 = \eta_2 = \eta_3 \end{cases} \quad (4.8)$$

Similarly, we get contradictions if η_1 or η_2 or $\eta_3 < \beta_1$ and if η_1 and η_2 and $\eta_3 > \beta_1$.

Then discuss the left cases as follows:

Case 1: $\alpha_1 = \theta_1 < \theta_2 \leq \theta_3$ and $\beta_1 = \eta_1 < \eta_2 \leq \eta_3$.

That follows $\gamma_1 C_1 = \delta_1 K_1$. $\alpha_1 = \theta_1$ and $\beta_1 = \eta_1$ implies $C_1 = K_1$, thus $\gamma_1 = \delta_1$.

Next, subtract the first term of both side, then let

$$f'(x) = f(x) - \gamma_1 C_1 x^{\alpha_1-1} (1-x)^{\beta_1-1} \text{ and}$$

$$g'(x) = g(x) - \delta_1 K_1 x^{\theta_1-1} (1-x)^{\eta_1-1}.$$

Note: the $f'(x)$ and $g'(x)$ does not represent the derivates of $f(x)$ and $g(x)$.

Use the similar method, we could get

$$\lim_{x \rightarrow 0} x^{1-\alpha_2} f'(x) = \lim_{x \rightarrow 0} x^{1-\alpha_2} g'(x) \quad (4.9)$$

then have

$$\gamma_2 C_2 = \begin{cases} \infty & \text{if } \theta_2 \text{ or } \theta_3 < \alpha_2 \\ 0 & \text{if } \theta_2 \text{ and } \theta_3 > \alpha_2 \\ \delta_2 K_2 & \text{if } \alpha_2 = \theta_2 < \theta_3 \\ \delta_2 K_2 + \delta_3 K_3 & \text{if } \alpha_2 = \theta_2 = \theta_3 \end{cases} \quad (4.10)$$

and

$$\lim_{x \rightarrow 1} (1-x)^{1-\beta_2} f'(x) = \lim_{x \rightarrow 1} (1-x)^{1-\beta_2} g'(x) \quad (4.11)$$

then

$$\gamma_2 C_2 = \begin{cases} \infty & \text{if } \eta_2 \text{ or } \eta_3 < \beta_2 \\ 0 & \text{if } \eta_2 \text{ and } \eta_3 > \beta_2 \\ \delta_2 K_2 & \text{if } \beta_2 = \eta_2 < \eta_3 \\ \delta_2 K_2 + \delta_3 K_3 & \text{if } \beta_2 = \eta_2 = \eta_3 \end{cases} \quad (4.12)$$

Case 1a: $\alpha_2 = \theta_2 < \theta_3$ and $\beta_2 = \eta_2 < \eta_3$ implies $\gamma_2 C_2 = \delta_2 K_2$, also $C_2 = K_2$, then we can conclude $\gamma_2 = \delta_2$.

Case 1b: If $\alpha_2 = \theta_2 = \theta_3$ and $\beta_2 = \eta_2 = \eta_3$, then $\gamma_2 C_2 = \delta_2 K_2 + \delta_3 K_3$. We can also have $C_2 = K_2 = K_3$, then implies $\gamma_2 = \delta_2 + \delta_3$. Hence, we get a contradiction as $\gamma_3 C_3 = 0$ in this case.

Finally, subtract the second term of each side, and let $f'' = \gamma_3 C_3 x^{\alpha_3-1} (1-x)^{\beta_3-1}$ and $g'' = \delta_3 K_3 x^{\theta_3-1} (1-x)^{\eta_3-1}$.

Construct the limit

$$\lim_{x \rightarrow 0} x^{1-\alpha_3} f''(x) = \lim_{x \rightarrow 0} x^{1-\alpha_3} g''(x) \quad (4.13)$$

this implies

$$\gamma_3 C_3 = \begin{cases} \infty & \text{if } \theta_3 < \alpha_3 \\ 0 & \text{if } \theta_3 > \alpha_3 \\ \delta_3 K_3 & \text{if } \theta_3 = \alpha_3 \end{cases} \quad (4.14)$$

and the limit

$$\lim_{x \rightarrow 0} (1-x)^{1-\beta_3} f''(x) = \lim_{x \rightarrow 0} (1-x)^{1-\beta_3} g''(x) \quad (4.15)$$

show that

$$\gamma_3 C_3 = \begin{cases} \infty & \text{if } \eta_3 < \beta_3 \\ 0 & \text{if } \eta_3 > \beta_3 \\ \delta_3 K_3 & \text{if } \eta_3 = \beta_3 \end{cases} \quad (4.16)$$

$\theta_3 = \alpha_3$ and $\eta_3 = \beta_3$ follows $\gamma_3 C_3 = \delta_3 K_3$, and we can see $C_3 = K_3$, thus $\delta_3 = \gamma_3$.

Case 2: $\alpha_1 = \theta_1 = \theta_2 < \theta_3$ and $\beta_1 = \eta_1 = \eta_2 < \eta_3$

That follows $\gamma_1 C_1 = \delta_1 K_1 + \delta_2 K_2$ and as $\alpha_1 = \theta_1 = \theta_2$ and $\beta_1 = \eta_1 = \eta_2$,

$C_1 = K_1 = K_2$. Then $\gamma_1 = \delta_1 + \delta_2$

Use the same approach above, subtract the first terms of $f(x)$ and the first two terms of $g(x)$ then construct limits:

$$\lim_{x \rightarrow 0} x^{1-\alpha_2} f'(x) = \lim_{x \rightarrow 0} x^{1-\alpha_2} g''(x) \quad (4.17)$$

and

$$\lim_{x \rightarrow 0} (1-x)^{1-\beta_2} f'(x) = \lim_{x \rightarrow 0} (1-x)^{1-\beta_2} g''(x) \quad (4.18)$$

which implies

$$\gamma_2 C_2 = \begin{cases} \infty & \text{if } \theta_3 < \alpha_2 \\ 0 & \text{if } \theta_3 > \alpha_2 \\ \delta_3 K_3 & \text{if } \theta_3 = \alpha_2 \end{cases} \quad (4.19)$$

and

$$\gamma_2 C_2 = \begin{cases} \infty & \text{if } \eta_3 < \beta_2 \\ 0 & \text{if } \eta_3 > \beta_2 \\ \delta_3 K_3 & \text{if } \eta_3 = \beta_2 \end{cases} \quad (4.20)$$

2.18 and 2.19 show that if $\theta_3 = \alpha_2$ and $\eta_3 = \beta_2$, $\gamma_2 C_2 = \delta_3 K_3$, as we have shown $\gamma_1 C_1 = \delta_1 K_1 + \delta_2 K_2$, we can conclude $\gamma_3 C_3 = 0$, here is a contradiction.

Case 3: $\alpha_1 = \theta_1 = \theta_2 = \theta_3$ and $\beta_1 = \eta_1 = \eta_2 = \eta_3$

That follows $\gamma_1 C_1 = \delta_1 K_1 + \delta_2 K_2 + \delta_3 K_3$ and as $\alpha_1 = \theta_1 = \theta_2 = \theta_3$ and $\beta_1 = \eta_1 = \eta_2 = \eta_3$, $C_1 = K_1 = K_2 = K_3$. Then $\gamma_1 = \delta_1 + \delta_2 + \delta_3$, hence $\gamma_2 C_2 + \gamma_3 C_3 = 0$, here is a contradiction.

Besides Case 1, Case 2 and Case 3, we also consider if $\delta_2 = \delta_3 = 0$, then $\gamma_1 C_1 = \delta_1 K_1$, as $\alpha_1 = \theta_1$ and $\beta_1 = \eta_1$, we have $\gamma_1 = \delta_1$, then we get $\gamma_2 C_2 + \gamma_3 C_3 = 0$, here is a contradiction.

Similarly, we could prove there exist contradictions if $\delta_1 = \delta_2 = 0$, $\delta_1 = \delta_3 = 0$, $\delta_1 = 0$, $\delta_2 = 0$ or $\delta_3 = 0$.

■

4.2.2 Identifiability of Three-component Contaminated Beta Model

Then, let us consider a three-component contaminated beta model:

$$(1 - \gamma_1 - \gamma_2)Beta(1, 1) + \gamma_1 Beta(\alpha_1, \beta_1) + \gamma_2 Beta(\alpha_2, \beta_2) \quad (4.21)$$

Let

$$f(x) = (1 - \gamma_1 - \gamma_2)Beta(1, 1) + \gamma_1 B(\alpha_1, \beta_1)x^{\alpha_1-1}(1-x)^{\beta_1-1} + \gamma_2 B(\alpha_2, \beta_2)x^{\alpha_2-1}(1-x)^{\beta_2-1} \quad (4.22)$$

and

$$g(x) = (1 - \delta_1 - \delta_2)Beta(1, 1) + \delta_1 B(\theta_1, \eta_1)x^{\theta_1-1}(1-x)^{\eta_1-1} + \delta_2 B(\theta_2, \eta_2)x^{\theta_2-1}(1-x)^{\eta_2-1} \quad (4.23)$$

We also could add some constraints to guarantee the identifiability of this model as follows:

Theorem 4.2.2. *A 3-component contaminated Beta mixture model (CB model) shown in 4.22 with γ_1 and γ_2 and $1 - \gamma_1 - \gamma_2 \in (0, 1)$, $0 < \alpha_1 < \alpha_2 < 1$ and $1 < \beta_1 < \beta_2$ cannot be expressed as a 3-component CB model shown as 4.23 with δ_1 and $\delta_2 \in [0, 1]$, $0 < \theta_1 \leq \theta_2 \leq 1$ and $1 \leq \eta_1 \leq \eta_2$ unless $\gamma_1 = \delta_1$, $\alpha_1 = \theta_1$, $\beta_1 = \eta_1$, $\gamma_2 = \delta_2$, $\alpha_2 = \theta_2$, $\beta_2 = \eta_2$.*

Proof:

Suppose $f(x) = g(x)$ and let $B(\alpha_i, \beta_i) = C_i$ and $B(\theta_j, \eta_j) = K_j$.

Assume that $0 < \alpha_1 < \alpha_2 < 1$ and $1 < \beta_1 < \beta_2, 0 < \theta_1 \leq \theta_2 \leq 1$ and $1 \leq \eta_1 \leq \eta_2$.

First of all, we can see the limit $\lim_{x \rightarrow 1} f(x) = \lim_{x \rightarrow 1} g(x)$ implies $(1 - \gamma_1 - \gamma_2) = (1 - \delta_1 - \delta_2)$ when $\theta_1 < \theta_2 \leq 1$ and $1 < \eta_1 < \eta_2$, or $(1 - \gamma_1 - \gamma_2) = (1 - \delta_1 - \delta_2) + \delta_1 \theta_1$ when $\theta_1 < \theta_2 \leq 1$ and $1 = \eta_1 < \eta_2$.

Next, if $\theta_1 < \theta_2 \leq 1$ and $1 < \eta_1 < \eta_2$, subtract the first term of both side, then let

$$f'(x) = \gamma_1 B(\alpha_1, \beta_1) x^{\alpha_1-1} (1-x)^{\beta_1-1} + \gamma_2 B(\alpha_2, \beta_2) x^{\alpha_2-1} (1-x)^{\beta_2-1}$$

and

$$g'(x) = \delta_1 B(\theta_1, \eta_1) x^{\theta_1-1} (1-x)^{\eta_1-1} + \delta_2 B(\theta_2, \eta_2) x^{\theta_2-1} (1-x)^{\eta_2-1}.$$

As we did in last proof, we could get

$$\lim_{x \rightarrow 0} x^{1-\alpha_1} f'(x) = \lim_{x \rightarrow 0} x^{1-\alpha_1} g'(x) \quad (4.24)$$

then have

$$\gamma_1 C_1 = \begin{cases} 0 & \text{if } \theta_2 \geq \theta_1 > \alpha_1 \\ \delta_1 K_1 & \text{if } \alpha_1 = \theta_1 < \theta_2 \\ \delta_1 K_1 + \delta_2 K_2 & \text{if } \alpha_1 = \theta_1 = \theta_2 \\ \infty & \text{otherwise} \end{cases} \quad (4.25)$$

and

$$\lim_{x \rightarrow 1} (1-x)^{1-\beta_1} f'(x) = \lim_{x \rightarrow 1} (1-x)^{1-\beta_1} g'(x) \quad (4.26)$$

then

$$\gamma_1 C_1 = \begin{cases} 0 & \text{if } \eta_2 \geq \eta_1 > \beta_1 \\ \delta_1 K_1 & \text{if } \beta_1 = \eta_1 < \eta_2 \\ \delta_1 K_1 + \delta_2 K_2 & \text{if } \beta_1 = \eta_1 = \eta_2 \\ \infty & \text{otherwise} \end{cases} \quad (4.27)$$

Since $\gamma_1 C_1 = 0$ or $\gamma_1 C_1 = \infty$ or $\gamma_1 C_1 = \delta_1 K_1 + \delta_2 K_2$ will make contradictions, we conclude $\gamma_1 C_1 = \delta_1 K_1$. And as $\alpha_1 = \theta_1$ and $\beta_1 = \eta_1$, $C_1 = K_1$. Then we can say $\gamma_1 = \delta_1$.

Repeat the step above, we can also prove $\alpha_2 = \theta_2$, $\beta_2 = \eta_2$, follows $\gamma_2 = \delta_2$.

If $\theta_1 < \theta_2 \leq 1$ and $1 = \eta_1 < \eta_2$, subtract the first term of $f(x)$, then let $g''(x) = g'(x) - \delta_1 \theta_1$. Take limitation as above,

$$\lim_{x \rightarrow 0} x^{1-\alpha_1} f'(x) = \lim_{x \rightarrow 0} x^{1-\alpha_1} g'(x) = \lim_{x \rightarrow 0} x^{1-\alpha_1} g''(x) \quad (4.28)$$

we get same result as 4.25, and

$$\lim_{x \rightarrow 1} (1-x)^{1-\beta_1} f'(x) = \lim_{x \rightarrow 1} (1-x)^{1-\beta_1} g''(x) \quad (4.29)$$

The right side does not exist unless $\beta_1 = 1$, and we get same result as 4.27. Repeat the previous steps, we get $\alpha_1 = \theta_1$, $\beta_1 = \eta_1 = 1$, $\gamma_1 = \delta_1$, $\alpha_2 = \theta_2$, $\beta_2 = \eta_2$, $\gamma_2 = \delta_2$.

If $\theta_1 = \theta_2 = 1$ and $\eta_1 = \eta_2 = 1$, then $g(x) = \text{Beta}(1,1)$, $f(x) = g(x)$ implies

$$\gamma_1 B(\alpha_1, \beta_1) x^{\alpha_1-1} (1-x)^{\beta_1-1} + \gamma_2 B(\alpha_2, \beta_2) x^{\alpha_2-1} (1-x)^{\beta_2-1} = (\gamma_1 + \gamma_2) \text{Beta}(1, 1)$$

but we could not found α and β with γ_1 and γ_2 and $1 - \gamma_1 - \gamma_2 \in (0, 1)$, $0 < \alpha_1 < \alpha_2 < 1$ and $1 < \beta_1 < \beta_2$.

similarly, we also check the condition, if $\theta_1 = \theta_2 < 1$ and $\eta_1 = \eta_2 > 1$, the equation $f(x) = g(x)$ could not be satisfied.

■.

4.2.3 Identifiability of Two-component Beta Mixture model

Now, let us consider a two-component beta mixture model:

Theorem 4.2.3. *A 2-component Beta mixture model with $\gamma_1, \gamma_2 \in (0, 1)$ and $\gamma_1 + \gamma_2 = 1$, $0 < \alpha_1 < \alpha_2$ and $0 < \beta_1 < \beta_2$ cannot be expressed as a 2-component beta mixture model with $\delta_1, \delta_2 \in [0, 1]$, $\delta_1 + \delta_2 = 1$, $0 < \theta_1 \leq \theta_2$, $0 < \eta_1 \leq \eta_2$ unless $\gamma_1 = \delta_1$, $\gamma_2 = \delta_2$, $\alpha_1 = \theta_1$, $\beta_1 = \eta_1$, $\gamma_2 = \delta_2$, $\alpha_2 = \theta_2$, $\beta_2 = \eta_2$.*

Proof:

Let

$$f(x) = \gamma_1 B(\alpha_1, \beta_1) x^{\alpha_1-1} (1-x)^{\beta_1-1} + \gamma_2 B(\alpha_2, \beta_2) x^{\alpha_2-1} (1-x)^{\beta_2-1} \quad (4.30)$$

and

$$g(x) = \delta_1 B(\theta_1, \eta_1) x^{\theta_1-1} (1-x)^{\eta_1-1} + \delta_2 B(\theta_2, \eta_2) x^{\theta_2-1} (1-x)^{\eta_2-1} \quad (4.31)$$

Suppose $f(x) = g(x)$ and let $B(\alpha_i, \beta_i) = C_i$ and $B(\theta_j, \eta_j) = K_j$. Assume that $\alpha_1 < \alpha_2$, $\beta_1 < \beta_2$, $\theta_1 \leq \theta_2$.

Multiply $f(x) = g(x)$ by $x^{1-\alpha_1}$ and take limits of both sides, we have

$$\lim_{x \rightarrow 0} x^{1-\alpha_1} f(x) = \lim_{x \rightarrow 0} x^{1-\alpha_1} g(x) \quad (4.32)$$

implies

$$\gamma_1 C_1 = \begin{cases} \infty & \text{if } \theta_1 \leq \theta_2 < \alpha_1 \\ 0 & \text{if } \theta_2 \geq \theta_1 > \alpha_1 \\ \delta_1 K_1 & \text{if } \alpha_1 = \theta_1 < \theta_2 \\ \delta_1 K_1 + \delta_2 K_2 & \text{if } \alpha_1 = \theta_1 = \theta_2 \end{cases} \quad (4.33)$$

multiply the equation $f(x) = g(x)$ by $(1-x)^{1-\beta_1}$ and take limits of both sides, we have

$$\lim_{x \rightarrow 1} (1-x)^{1-\beta_1} f(x) = \lim_{x \rightarrow 1} (1-x)^{1-\beta_1} g(x) \quad (4.34)$$

implies

$$\gamma_1 C_1 = \begin{cases} \infty & \text{if } \eta_1 \text{ or } \eta_2 < \beta_1 \\ 0 & \text{if } \eta_1 \text{ and } \eta_2 > \beta_1 \\ \delta_1 K_1 & \text{if } \beta_1 = \eta_1 < \eta_2 \\ \delta_1 K_1 + \delta_2 K_2 & \text{if } \beta_1 = \eta_1 = \eta_2 \end{cases} \quad (4.35)$$

Case1 if $\alpha_1 = \theta_1 < \theta_2$ and $\beta_1 = \eta_1 < \eta_2$, $\gamma_1 C_1 = \delta_1 K_1$, we can get $C_1 = K_1$, then $\gamma_1 = \delta_1$.

Then subtract the first terms and do the same thing, we can get $\alpha_2 = \theta_2$, $\beta_2 = \eta_2$, then $\gamma_2 = \delta_2$.

Case2 if $\alpha_1 = \theta_1 = \theta_2$ and $\beta_1 = \eta_1 = \eta_2$, we have $\gamma_1 C_1 = \delta_1 K_1 + \delta_2 K_2$. Here is a contradiction, because that implies $\gamma_2 C_2 = 0$.

Case3 if $\alpha_1 = \theta_1 = \theta_2$ and $\beta_1 = \eta_1 < \eta_2$, then we have $\gamma_1 C_1 = \delta_1 K_1 = \delta_1 K_1 + \delta_2 K_2$, which implies $\delta_2 K_2 = 0$ and $\gamma_2 C_2 = 0$, here is the contradiction.

Case4 if $\alpha_1 = \theta_1 < \theta_2$ and $\beta_1 = \eta_1 = \eta_2$, similarly as case 3, we have a contradiction.

■

Now we consider:

H_0 : P-values are distributed as a two-component contaminated beta mixture model.

H_1 : P-values are distributed as a three-component contaminated beta mixture model.

Let' consider the Likelihood Ratio Test, log-likelihood function of these models are

$$l_n(\pi, \alpha, \beta) = \sum_{i=1}^n \log[(1 - \pi)f(X_i; 1, 1) + \pi f(X_i; \alpha, \beta)] \quad (4.36)$$

and

$$\begin{aligned}
& l_n(\pi_1, \pi_2, \alpha_1, \beta_1, \alpha_2, \beta_2) \\
&= \sum_{i=1}^n \log[(1 - \pi_1 - \pi_2)f(X_i; 1, 1) + \pi_1 f(X_i; \alpha_1, \beta_1) + \pi_2 f(X_i; \alpha_2, \beta_2)]
\end{aligned} \tag{4.37}$$

Use EM algorithm to get the estimate of parameters, then the LRT can be expressed as

$$T_n = 2l_n(\hat{\pi}_1, \hat{\pi}_2, \hat{\alpha}_1, \hat{\beta}_1, \hat{\alpha}_2, \hat{\beta}_2) - 2l_n(\hat{\pi}, \hat{\alpha}, \hat{\beta}) \tag{4.38}$$

4.3 Estimating the MLEs

As we need to compute the maximum of likelihood estimates of the Beta contamination model, do constraint-optimization in multidimensional space in this section if we want to obtain the MLEs.

We first consider BFGS method as we did before in the Chapter 2, but as we have $0 < \alpha_1 < \alpha_2 < 1$ and $1 < \beta_1 < \beta_2$, the function "optim" we used in Chapter 2 could not deal with multidimensional optimization problem with more than one inequality.

Then we consider a parameter transformation method.

If we transform the parameter to

$$\begin{aligned}
\alpha_2 &= e^{-e^{u_2}} \\
\alpha_1 &= \alpha_2 * e^{-e^{u_1}} \\
\beta_2 &= e^{e^{v_2}} \\
\beta_1 &= \beta_2 * e^{e^{v_1}}
\end{aligned} \tag{4.39}$$

Then we get a new system of parameter of u and v,

$$\begin{aligned}
 u_1 &= \log\left(-\log\left(\frac{\alpha_1}{\alpha_2}\right)\right) \\
 u_2 &= \log(-\log(\alpha_2)) \\
 v_1 &= \log\left(\log\left(\frac{\beta_1}{\beta_2}\right)\right) \\
 v_2 &= \log(\log(\beta_2))
 \end{aligned}
 \tag{4.40}$$

With this transformation, the constraints $0 < \alpha_1 < \alpha_2 < 1$ and $1 < \beta_1 < \beta_2$ are satisfied, we could use function `optim` to do the constraint optimization to get the estimates of u_1, u_2 and v_1, v_2 . Then the MLEs are obtained by plug them in the formula 4.39.

After investigation, in some cases, the parameter transformation method tends to underestimate the weights of contamination fraction.

Then we consider another r function called "constrOptim" with Nelder–Mead method to do the multidimensional constrained optimization problem. This function could deal with multiple linear inequality constraints. The feasible region is defined by $u_i \theta - c_i \geq 0$, where u_i is constraint matrix and c_i is constraint vector, θ is the parameter vector.[R Core Team,]

So our constraints could be written as

$$\begin{bmatrix} -1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{bmatrix} \geq \begin{bmatrix} \epsilon \\ \epsilon \\ \epsilon \\ \epsilon \end{bmatrix}
 \tag{4.41}$$

Where ϵ is a very small number.

After investigation, if we set initial values properly, in 83% cases, the difference between the true value and the estimated maximum log-likelihood are smaller than 5%; in 17% cases, the difference are larger than 5% of the true value, but the value still not far from the true value(all difference are less than 10%). The time cost of the constrOptim function is half of the parameter transformation method.

4.4 Hypothesis Testing

Consider the hypothesis:

H_0 : P-values are distributed as a two-component contaminated beta mixture model.

H_1 : P-values are distributed as a three-component contaminated beta mixture model.

First, we considered to do some simulation study to check the null limiting distribution with constraints.

For each of several sample size, I generate 5000 data sets from the following null distribution

$$\begin{aligned}
 &0.8Beta(1, 1) + 0.2Beta(0.5, 1.5) \\
 &0.5Beta(1, 1) + 0.5Beta(0.5, 1.5) \\
 &0.2Beta(1, 1) + 0.8Beta(0.5, 1.5) \\
 &0.8Beta(1, 1) + 0.2Beta(0.25, 4) \\
 &0.5Beta(1, 1) + 0.5Beta(0.25, 4) \\
 &0.2Beta(1, 1) + 0.8Beta(0.25, 4)
 \end{aligned} \tag{4.42}$$

Calculate the LRT statistics. Figure 4.1 show the histogram of LRT statistic. If I plot the 25, 50, 75, 95, 99 percentile of the statistics vs. sample size in figure 4.2, 4.3, 4.4, we see it become more and more consistent as the sample size increase, and that

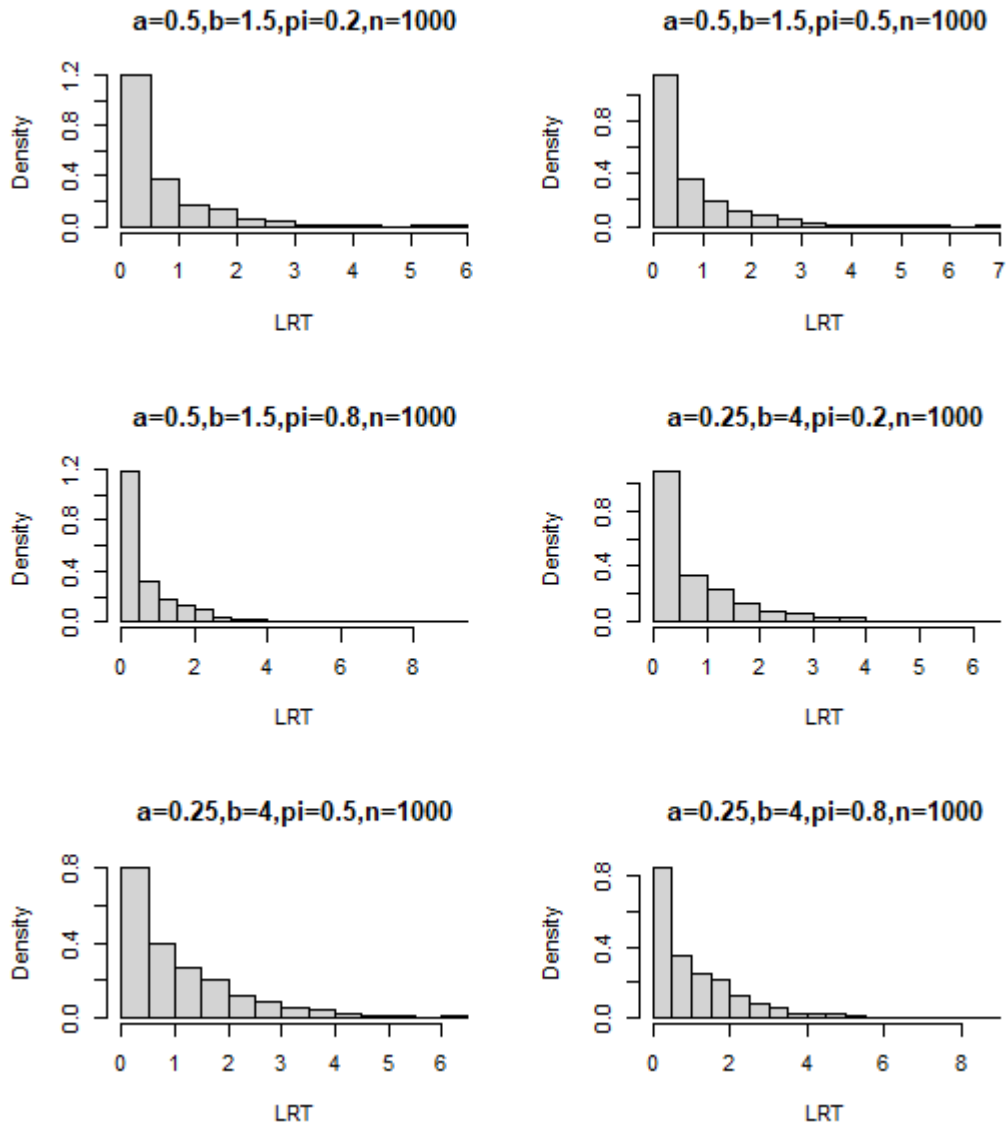


Figure 4.1: Histogram of LRT when sample size $n=1000$

indicates the null limiting distribution may exist.

With the constraints, getting an analytical expression of the null limiting distribution becomes complicated. So we consider a parametric bootstrapping to do the test. We have used this method in Chapter 3, section 3.4; the computing algorithm is similar.

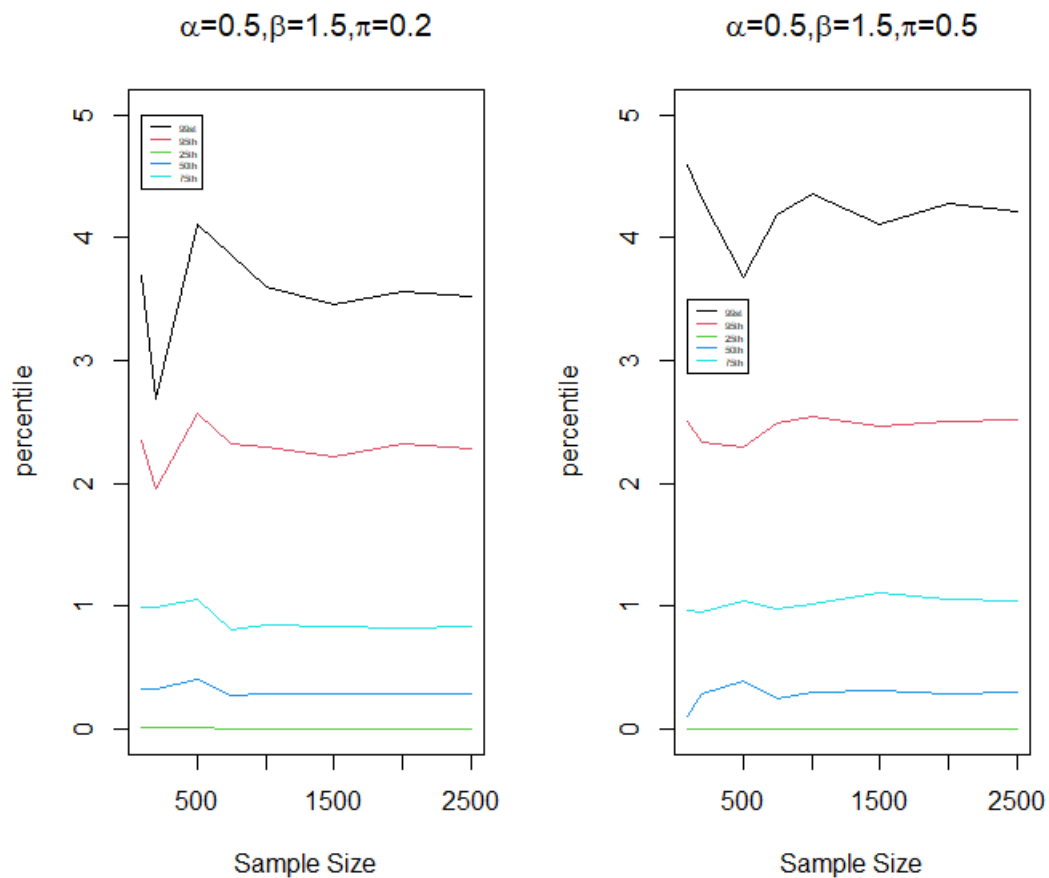


Figure 4.2: Percentile plots of LRT statistic under H_0 vs sample size

4.5 Introduction of sBIC

The disadvantage of the bootstrap method is if the sample size is huge, it is very time-consuming to do a sufficient number of repetitions. We also have to deal with a bunch of initial values. Thus we consider a model selection criteria: sBIC(singular Bayesian information criterion).

To continue the discussion in Chapter 1, it is known that both Akaike Information Criterion (AIC) or Bayesian information criterion (BIC) are not proper tools to solve model selection problems with singular issues, which come with Fisher information matrices that are not invertible. ([Keribin, 2000], [Drton et al., 2009]). Because of

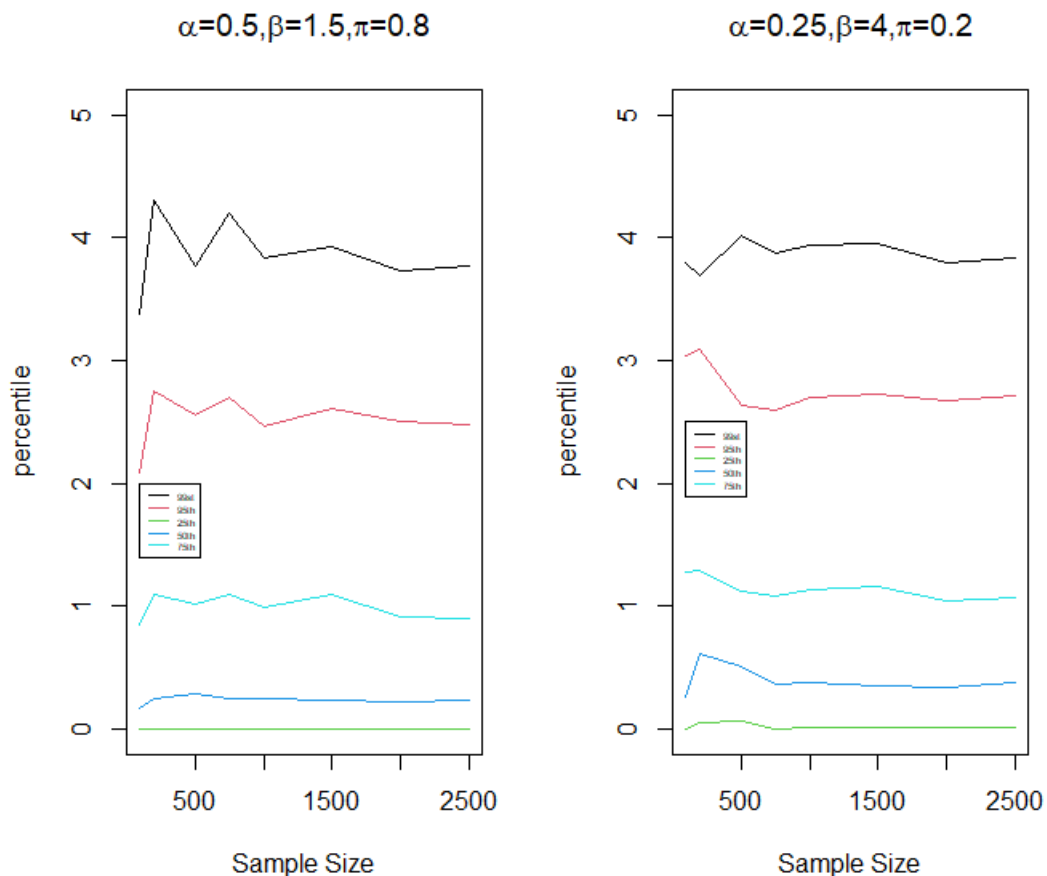


Figure 4.3: Percentile plots of LRT statistic under H_0 vs sample size

it, AIC and BIC should not be used for some mixture modeling applications. For example, to calculate the number of components, when three or more components are needed in the mixture model.

To circumvent this issue, a novel information criterion, singular Bayesian information criterion (sBIC), was introduced by [Drton and Plummer, 2017]. It is a modified Bayesian information criterion that handles singular model selection problems by addressing the invertibility issue of Fisher information matrices due to singularity. Let us take a closer look at the details of the sBIC following. All formulas and notations below are from [Drton and Plummer, 2017].

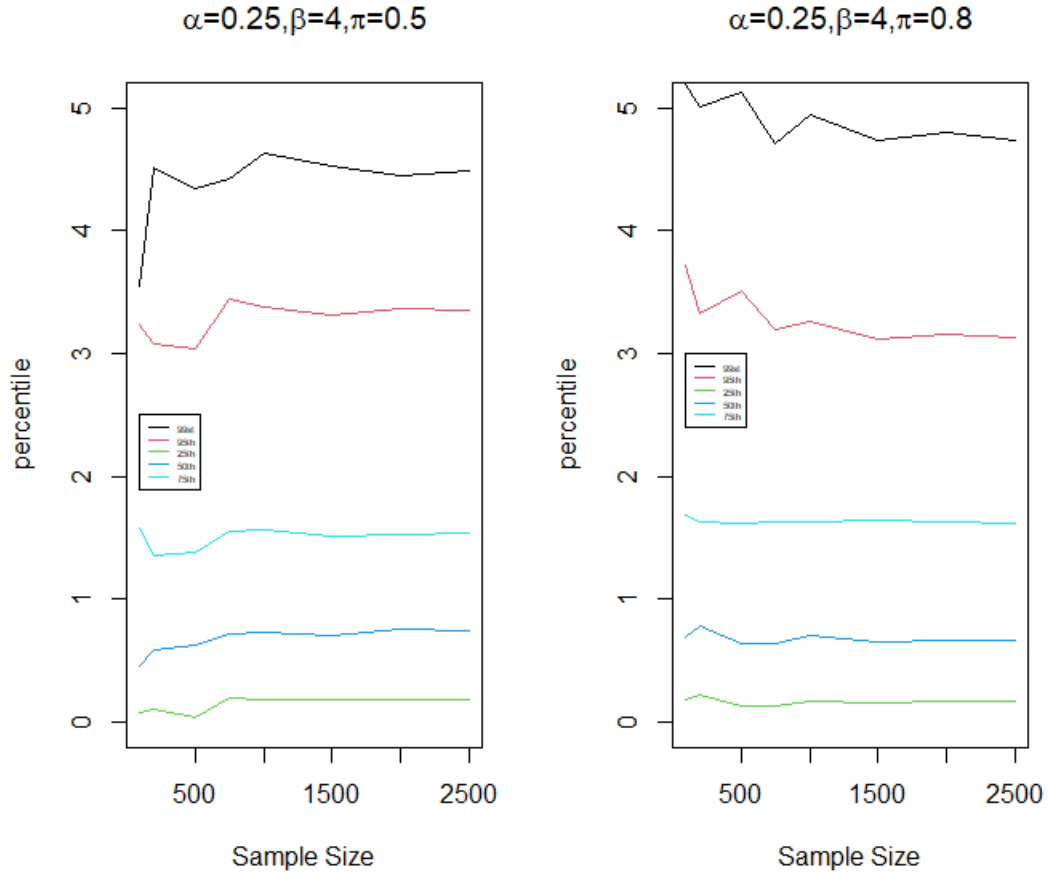


Figure 4.4: Percentile plots of LRT statistic under H_0 vs sample size

Let $Y_n = (Y_{n1}, Y_{n1}, \dots, Y_{nn})$ be n iid observations in a sample, $\{M_i : i \in I\}$ be a finite set of candidate models for the distribution of the observations.

$$L(M_i) := P(\mathbf{Y}_n | M_i) = \int_{M_i} P(\mathbf{Y}_n | \pi_i, M_i) dP(\pi_i | M_i) \quad (4.43)$$

where $P(\pi_i | M_i)$ is prior distribution for $\pi_i \in M_i$, $P(\mathbf{Y}_n | \pi_i, M_i)$ is the likelihood function.

Under suitable conditions, [Schwarz, 1978] observed

$$\log[L(M_i)] = \log[P(\mathbf{Y}_n | \hat{\pi}_i, M_i)] - \frac{d_i}{2} \log(n) + O_p(1) \quad (4.44)$$

Notation is in [Drton and Plummer, 2017]. Where the $P(\mathbf{Y}_n | \hat{\pi}_i, M_i)$ is the maximum of the likelihood function.

The resulting BIC for model M_i is

$$BIC(M_i) = \log[P(\mathbf{Y}_n | \hat{\pi}_i, M_i)] - \frac{d_i}{2} \log(n) \quad (4.45)$$

[Drton and Plummer, 2017] said that it is impossible to get a large sample quadratic approximation of a log-likelihood function for singular model. Formula 4.44 is false for singular model. [Watanabe, 2009] show the property of singular model for Y_n generated from $\pi_0 \in M_i$

$$\log[L(M_i)] = \log[P(\mathbf{Y}_n | \pi_0, M_i)] - \lambda_i(\pi_0) \log(n) + [m_i(\pi_0) - 1] \log[\log(n)] + O_p(1) \quad (4.46)$$

where $\lambda_i(\pi_0)$ is a rational number called learning coefficient and $m_i(\pi_0)$ is the multiplicity of the learning coefficient, it is an integer $\in \{1, 2, \dots, d_i\}$ where d_i is dimension of parameter space.

[Drton and Plummer, 2017] shows that the likelihood ratios are bounded in probability for exponential families, so the log-likelihood could be expressed in terms of maximum of the log-likelihood function:

$$\log[L(M_i)] = \log[P(\mathbf{Y}_n | \hat{\pi}_i, M_i)] - \lambda_i(\pi_0) \log(n) + [m_i(\pi_0) - 1] \log[\log(n)] + O_p(1) \quad (4.47)$$

Then [Drton and Plummer, 2017] define the sBIC as

$$sBIC(M_i) = \log[L'(M_i)] \quad (4.48)$$

where $L'(M_i)$ is unique solution of 4.49

$$\sum_{j \leq i} [L'(M_i) - L'_{ij}] L'(M_j) P(M_j) = 0, i \in I. \quad (4.49)$$

And L'_{ij} is

$$L'_{ij} = P(\mathbf{Y}_n \mid \hat{\pi}_i, M_i) \frac{(\log(n))^{m_{ij}-1}}{n^{\lambda_{ij}}} > 0 \quad (4.50)$$

Where λ_{ij} and m_{ij} are constant such that $\lambda_i(\pi_0) = \lambda_{ij}$ and $m_i(\pi_0) = m_{ij}$. Since for all singular model selection problem, the learning coefficient and multiplicity is almost surely constant.

[Drton and Plummer, 2017] also show the sBIC could be expressed in the form

$$sBIC(M_i) = \log[P(\mathbf{Y}_n \mid \hat{\pi}_i, M_i)] - \text{penalty}(M_i), \quad (4.51)$$

where $\text{penalty}(M_i) \leq \frac{1}{2} \dim(M_i) \log(n)$.

In our study, M_i is CB model with i mixture components. There are two candidate models, two component CB model and three component CB model.

According to [Watanabe, 2009] and [Drton and Plummer, 2017], $m_{ij} = 1$ we have

$$\lambda_{ij} \leq \frac{1}{2} [\dim(M_i) - 2(i - j)] = \frac{1}{2} [3i - 3 - 2(i - j)] = \frac{1}{2} (i + 2j - 3)$$

where i is the number of mixture components in the learning machine and j is the number of components in a true model ($j \leq i$).

4.6 Simulation

We generated data sets from the following null distributions for different sample sizes

4.6.1 Null distribution

$$\begin{aligned}(N31) & 0.3\text{Beta}(1, 1) + 0.7\text{Beta}(0.5, 1.5) \\(N32) & 0.5\text{Beta}(1, 1) + 0.5\text{Beta}(0.2, 5) \\(N33) & 0.7\text{Beta}(1, 1) + 0.3\text{Beta}(0.5, 1.5) \\(N34) & 0.7\text{Beta}(1, 1) + 0.3\text{Beta}(0.7, 2)\end{aligned}\tag{4.52}$$

4.6.2 Alternative distributions

$$\begin{aligned}(A31) & 0.4\text{Beta}(1, 1) + 0.3\text{Beta}(0.3, 1.5) + 0.3\text{Beta}(0.7, 6) \\(A32) & 0.6\text{Beta}(1, 1) + 0.2\text{Beta}(0.3, 1.5) + 0.2\text{Beta}(0.7, 6) \\(A33) & 0.6\text{Beta}(1, 1) + 0.3\text{Beta}(0.3, 1.5) + 0.1\text{Beta}(0.7, 6) \\(A34) & 0.6\text{Beta}(1, 1) + 0.2\text{Beta}(0.3, 2) + 0.2\text{Beta}(0.8, 4) \\(A35) & 0.6\text{Beta}(1, 1) + 0.2\text{Beta}(0.4, 1.5) + 0.2\text{Beta}(0.6, 6) \\(A36) & 0.8\text{Beta}(1, 1) + 0.1\text{Beta}(0.2, 2) + 0.1\text{Beta}(0.5, 4)\end{aligned}\tag{4.53}$$

We performed 2000 repetitions in the simulations and set the significance level as 0.05. Since the data sets are simulated, we use the true value as the initial point when fitting the mixture model to obtain the maximum log-likelihood function, and all the initial weights are equal. That makes the convergence time of the EM algorithm to be shorter. I use bootstrapping method and sBIC. Simulation results are shown in Figure 4.5 and 4.6.

When we look at Figure 4.5, the actual rejection rate is how many times out of 2000 we reject the H_0 based on bootstrapping. The actual rejection rates of the four null distributions are not far from 0.05.

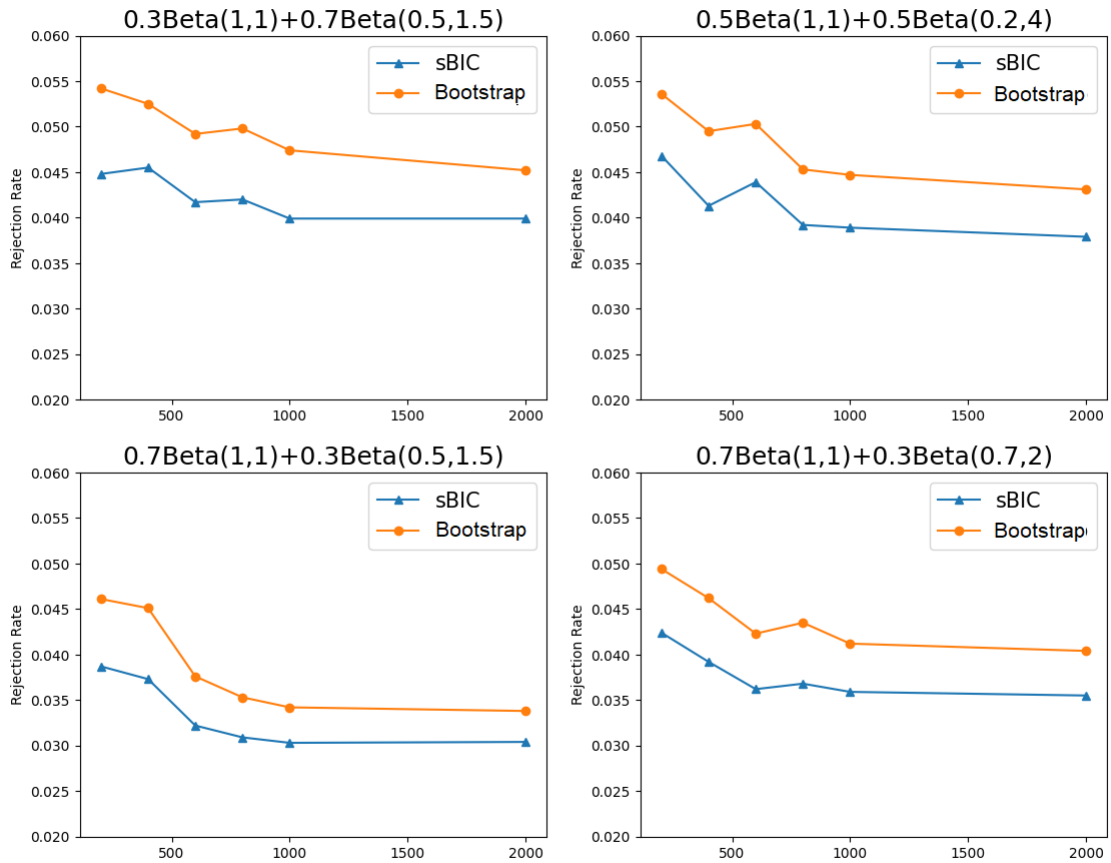


Figure 4.5: Actual rejection rate

Power curves are shown in Figure 4.6; our concerns are the performance of the test under the alternative. As shown in the figure, the bootstrap method works better than the sBIC in general, but the time cost of the bootstrap method is much longer, especially when the sample size and the bootstrap size are getting larger. For this simulation study, the sBIC method cost several hours, but the bootstrap method cost over a month. The power of scenario (A31) is higher than (A32), which is reasonable because as the mixing proportion of Beta(1,1) increases, it is harder to distinguish the two contaminations. The power of scenario (A33) is small than (A32). It is plausible

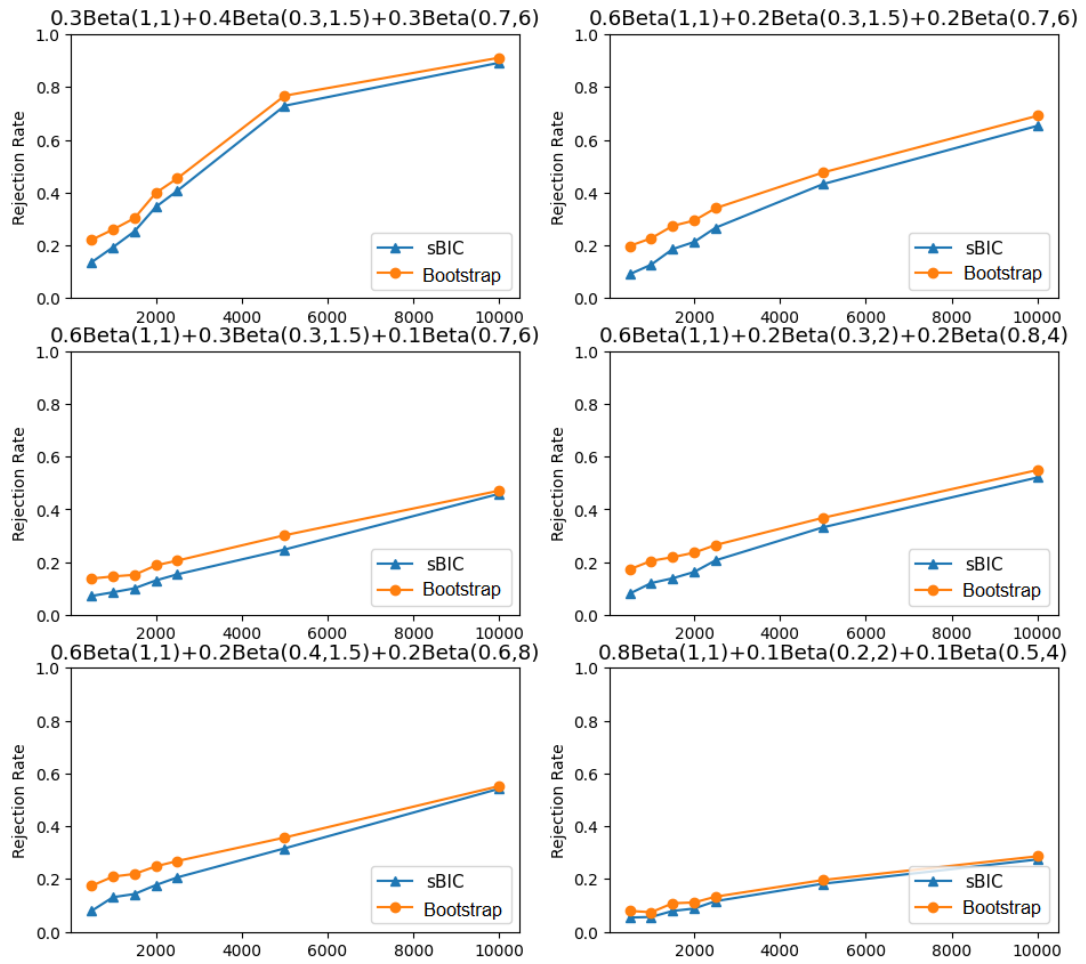


Figure 4.6: Power curves

since the mixing proportion of contamination of (A32) is equal(0.2 versus 0.2), while for the (A33), the mixing proportions are 0.3 versus 0.1. When the contamination fraction is quite small in (A36), the power turns out to be quite low. The power of (A32) is higher than (A34) and (A35) since the two shape parameter value has a bigger gap between the two contaminations in (A32).

4.7 Real Data Application

4.7.1 Introduction of real data

We continue to the microarray data we used in previous chapters, which is data on the systematic genome-wide DNA methylation alternation in blood cells of toddlers with Down Syndrome. 34 children with age 0.5–4.5 years take part in this study, 17 has Down syndrome, and 17 are typically developing children[Naumova et al., 2021]. Data 1, Data 2 and Data 3 are same with Chapter 2. Introductions are in section 2.5.1.

Data 4: We randomly select 20 data sets without replacement with a sample size $n = 10000$ from Data 1.

4.7.2 Results

Data1:

We have fitted a two-component Beta Contamination model to data 1 in Chapter 1, the fitted model is

$$0.697Beta(1, 1) + 0.303Beta(0.363, 1.997) \quad (4.54)$$

We also fit a three-component Beta contamination model to data 1, and obtain a fitted model below

$$0.697Beta(1, 1) + 0.191Beta(0.350, 1.508) + 0.112Beta(0.683, 4.958) \quad (4.55)$$

The histogram is shown in Figure 4.7, the red line shows the two-component fitted model, and the green line shows the three-component fitted model.

Then we used two methods: the bootstrap method with a bootstrap size of 2000 and sBIC; the sBIC selected the 3 component model, and the bootstrap rejected the null. Therefore, both ways preferred the three-component Beta contamination model.

In formula 4.55, the second component has a more extreme α value and the third component has a more extreme β value, when I plot and compare their density, the density of second component has thicker tail than the third component. It suggest the second component involve more large p-values. In this case, we say the second component corresponds to moderate differentially expressed gene group and the third component corresponds to high differentially expressed gene group.

Then according to the fitted three-component Beta contamination model, and we have 461,258 methylations in the data, the estimate $\hat{\pi}_1 = 0.191$ and $\hat{\pi}_2 = 0.112$ indicated that about 88100 genes were moderately differentially expressed of the control group and Down syndrome group, about 51661 genes were highly differentially expressed of the control group and Down syndrome group, and about 319191 genes are not differentially expressed.

Data 2:

Histogram of Data 1

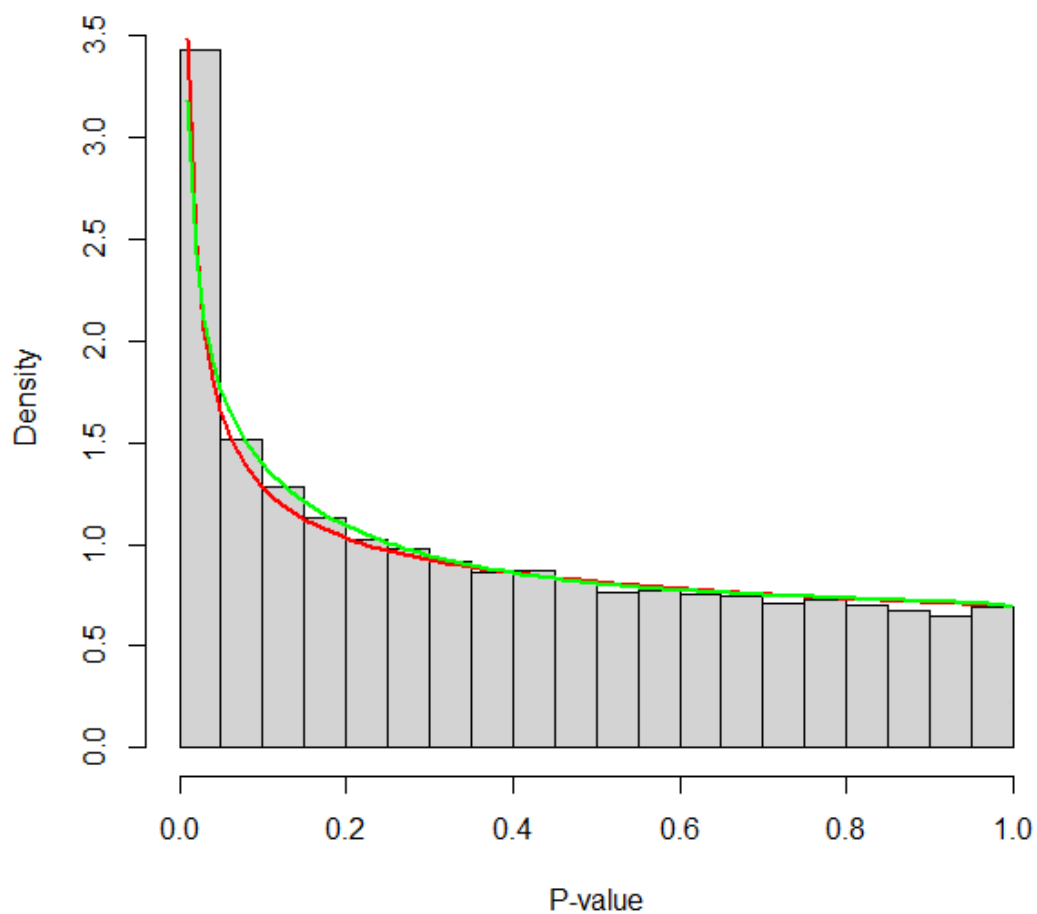


Figure 4.7: Histogram of Data 1 and fitted model, the red line shows the two-component fitted model, and the green line shows the three-component fitted model.

We also fitted a two-component Beta contamination model to the CHR21 data (Data 2), the fitted model is

$$0.425Beta(1, 1) + 0.575Beta(0.301, 3.135) \tag{4.56}$$

Histogram of p-values of CHR 21(Data2)

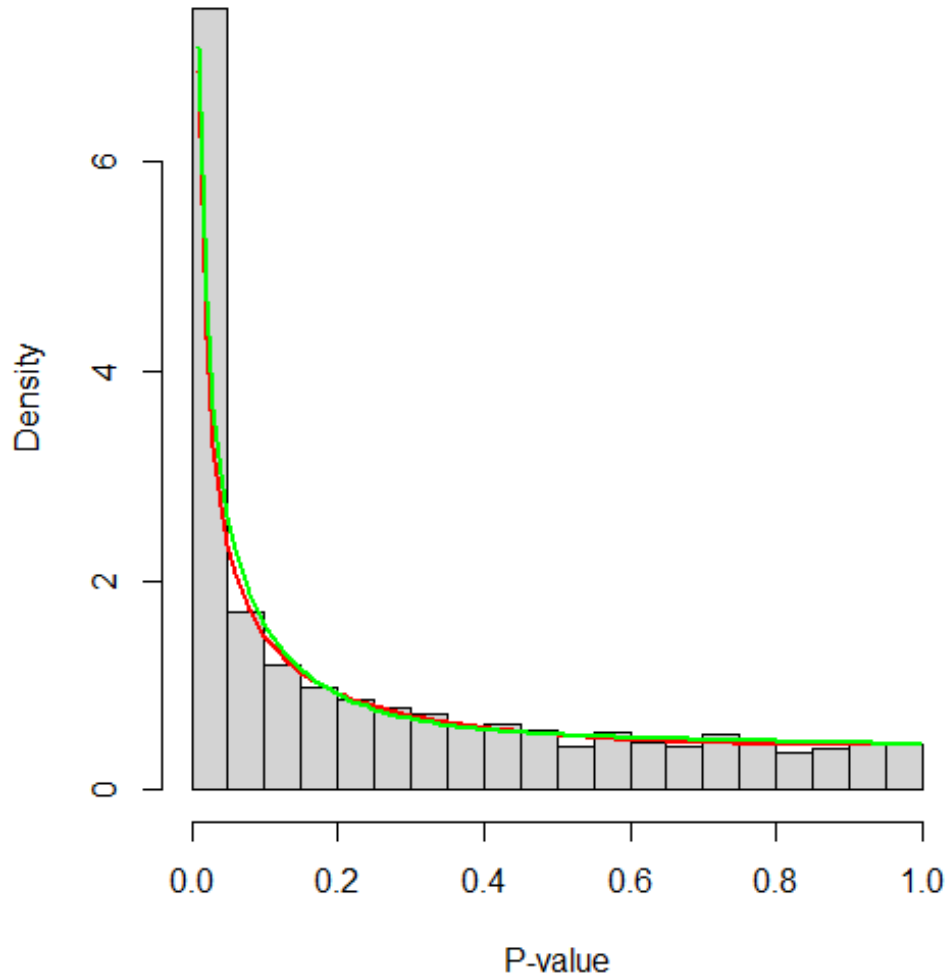


Figure 4.8: Histogram of Data 2 and fitted model, the red line shows the two-component fitted model, and the green line shows the three-component fitted model.

The fit a three-component Beta contamination model with constraints to data 2, the fitted model is

$$0.423Beta(1, 1) + 0.176Beta(0.253, 1.523) + 0.401Beta(0.504, 7.141) \quad (4.57)$$

The histogram is shown in Figure 4.8, the red line shows the two-component fitted

model, and the green line shows the three-component fitted model.

Then we used two methods: the bootstrap method with 2000 repetitions and sBIC. The sBIC selects the 2 component model with the 7% higher sBIC value than 3 component model; the bootstrap Fails to reject the null with p-value 0.18. Therefore, both methods indicate the p values follow a two-component Beta contamination model.

And according to the fitted two-component Beta contamination model, and we have 4205 methylations in data 2, the estimate $\hat{\pi} = 0.575$ indicated that the gene expression levels are different in about 2418 out of 4205 genes on Chromosome 21 between control group and Down syndrome group.

The result is reasonable because all genes of data 2 are located on Chromosome 21; for most of the differentially expressed genes, the difference in expression level may be more significant in the data. Or the moderate differentially expressed gene group is too small to get detected by test or sBIC.

Data 3:

We already fitted a two-component Beta contamination model to the data 3 in chapter 2, the fitted model is

$$0.705Beta(1, 1) + 0.295Beta(0.339, 1.834) \quad (4.58)$$

And the fitted constrained three-component Beta contamination model is:

$$0.705Beta(1, 1) + 0.187Beta(0.314, 1.104) + 0.108Beta(0.739, 5.848) \quad (4.59)$$

A histogram shown shows the fitted model in Figure 4.9, the red line shows the two-component fitted model, and the green line shows the three-component fitted model.

The same methods were applied to data 3, and the null hypothesis was rejected with both sBIC and bootstrapping methods.

According to the fitted three-component, Beta contamination model, and we have 452,477 methylations in the data 3, the estimate $\hat{\pi}_1 = 0.187$ and $\hat{\pi}_2 = 0.108$ indicated that about 84613 genes were moderately differentially expressed of the control group and Down syndrome group, about 48868 genes were highly differentially expressed of the control group and Down syndrome group, and about 318544 genes are not differentially expressed.

Data 4:

We fitted two-component and three-component models to the 20 data sets. The figure 4.7 shows the histograms and fitted models of the first 6 data, the red line is fitted constrained two-component Beta contamination model, and the green line shows the constrained three-component Beta contamination model. Table 4.1 list some MMLEs with constrained two-component and three-component Beta contamination model.

We use bootstrapping to do hypothesis tests and use sBIC to make a model selection. With the bootstrapping method, 14 of 20 rejected the null. And 13 data sets out of 20 selected three-component models via sBIC.

Histogram of Data3

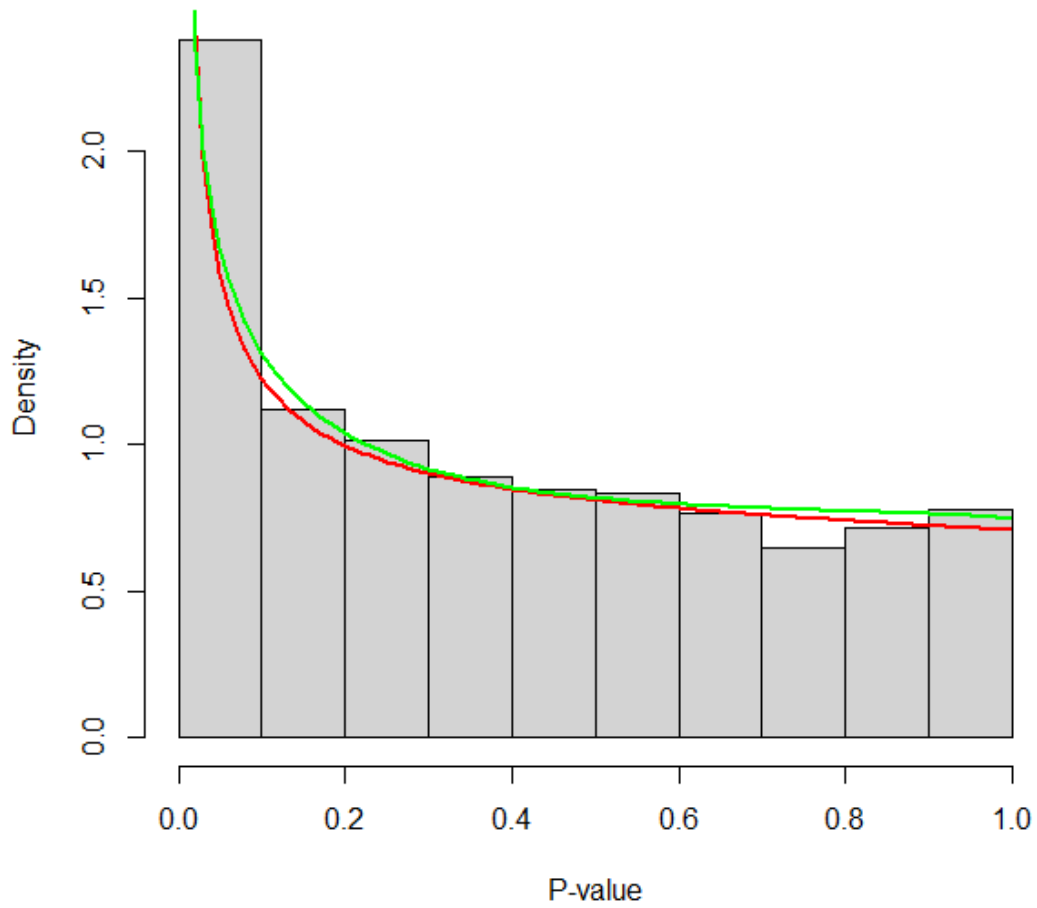


Figure 4.9: Histogram of Data 3 and fitted model, the red line shows the two-component fitted model, and the green line shows the three-component fitted model.

Table 4.1: MLEs of some fitted constrained model in data 4

data set	Two-component model			Three-component model					
	$\hat{\alpha}$	$\hat{\beta}$	π	α_1	β_1	$\hat{\pi}_1$	$\hat{\alpha}_2$	$\hat{\beta}_2$	$\hat{\pi}_2$
1	0.393	1.679	0.379	0.478	1.345	0.146	0.723	6.183	0.232
2	0.399	1.183	0.462	0.523	1.038	0.312	0.632	8.178	0.149
3	0.375	1.376	0.376	0.426	1.215	0.190	0.523	4.238	0.187
4	0.381	1.962	0.352	0.299	1.103	0.144	0.834	5.632	0.206
5	0.359	1.335	0.405	0.450	1.096	0.314	0.716	7.325	0.092
6	0.377	3.190	0.310	0.375	1.221	0.199	0.703	6.792	0.110

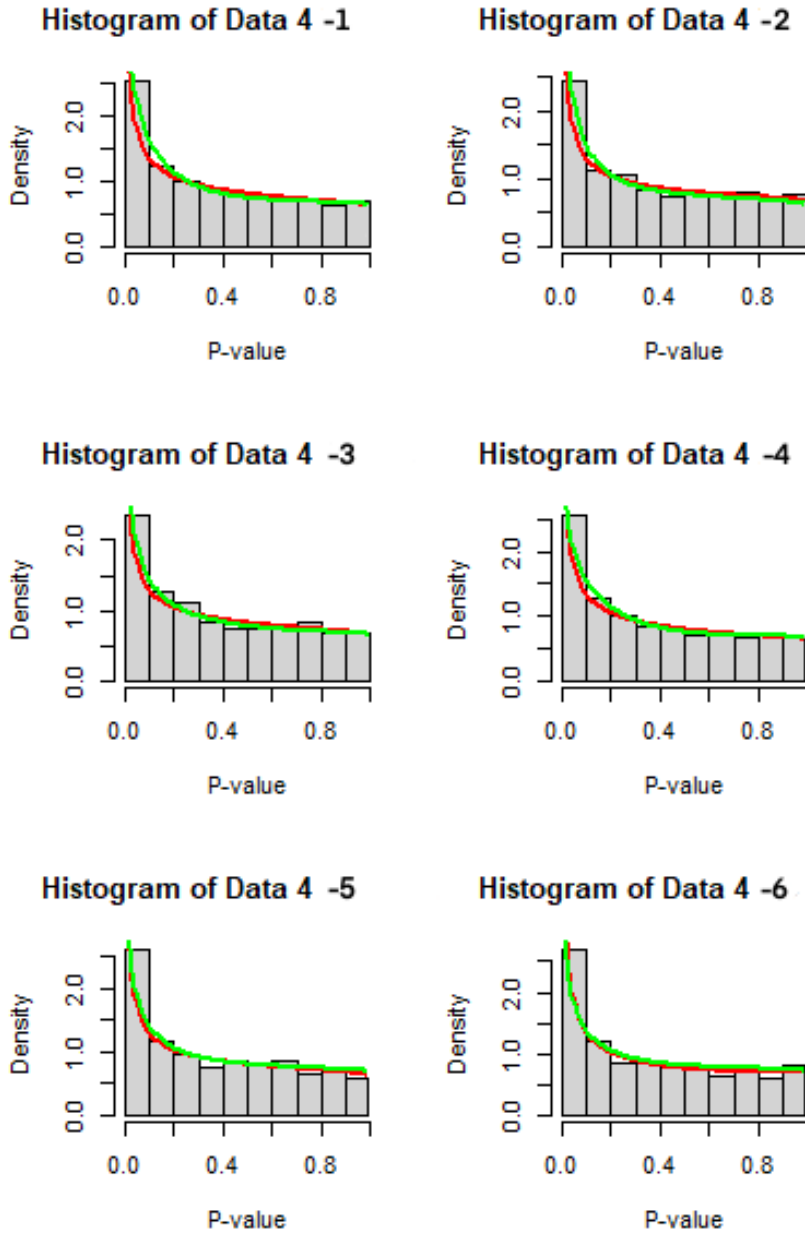


Figure 4.10: Histogram of Examples of Data 4 and fitted model, the red line shows the two-component fitted model, and the green line shows the three-component fitted model.

Chapter 5 Summary and future Work

My research is focused on the beta contamination model and its application to microarray data. In Chapter 2, we designed a constraints for two-component Beta contamination model to improve the power of modified likelihood ratio test(MLRT) [Chen et al., 2001][Dai and Charnigo, 2008]. For microarray data, if the null hypothesis is false, the distribution of the p-values is right-skewed and concentrated to zero. Thus, by designing a testing procedure with constraints to put the mode to the left end, the test would be more sensitive to the micro-array data. Using the actual critical value obtained by simulation might be suggestive when the sample size is small. We got simulated critical value as the subset of sample size, so I interpolated between points of sample size and critical value. The critical value is a decreasing function of sample size; with a bunch of points, we fit a linear model to determine the critical value in terms of sample size. We obtain critical values for different sample sizes without repetitive simulation.

A three-component Beta contamination model might better fit real data, especially when we are interested in distinguishing high differentially expressed genes and moderate differentially expressed genes in microarray data. Thus, in Chapters 3 and 4, we focused on testing the hypothesis of the component number of the Beta contamination model $g=2$ versus $g=3$.

Chapter 3 first considered a Beta contamination model with a kernel distribution from one parameter family by fixing the other shape parameter across all the components. The reason we consider the mixture model with one parameter family is the modified likelihood ratio test(MLRT) proposed by [Chen et al., 2004], has a

simple asymptotic null limiting distribution under some regularity conditions and is easy to apply. And fixing one shape parameter guarantees the identifiability of a three-component beta contamination model. We applied the MLRT and found some shortcomings after investigation. When we fix shape parameter α , a three-component beta contamination model gives a better fit than a two-component model even the true model is two component. With this model, we may make large type I error. On the other hand, the CB model fix the β parameter family seems have the risk of over-estimating the weights of two contamination. In chapter 4, we designed constraints for the three-component Beta contamination model and proved the identifiability of the model under these constraints. We used a likelihood-based test and bootstrap to test the hypothesis of the two-component vs. three contamination beta model. We also used a model selection criteria sBIC developed by [Drton and Plummer, 2017] to determine the number of components.

The contaminated Beta model and its application to microarray data have been explored extensively in previous research, and in this dissertation, some aspects remain for future research.

First, the likelihood-based test and sBIC need a large sample size to have appropriate power. If the sample size is at least 5000, the test and sBIC can be used comfortably. It is worth considering some procedures to improve the test power. It is not easy but one can still try some other ideas, such as designing another tighter constraint for the model.

Secondly, although we used bootstrap to accomplish the hypothesis test in chapter 4, it is worth doing some further study on the asymptotic properties of the null limiting distribution of the constrained Beta contamination model. Because the boot-

strap is complicated and time-consuming when the sample size is enormous, it is often true with the micro-array data. I used a supercomputer at the Chinese Academy of Science to do the simulation and real data analysis, and it still cost almost a month.

My committee member Dr. Derek Young gave valuable suggestion of considering proposing different scale constraints. For example, instead of using shape parameter $\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3$, one can reparametrize the parameter with scale coefficients, such as $\alpha, c_1\alpha, c_2\alpha, \beta, c_3\beta, c_4\beta$, where suitable values for the coefficients c_1, c_2, c_3, c_4 need to be set. Although we need additional constraints to the coefficients with this method, reducing the number of shape parameters involved in the three component beta model may mitigate difficulties when we try to obtaining the asymptotic properties of our log likelihood test.

Finally, we made independent assumptions when we did the research, but in microarray data, expression levels for each gene are correlated.[Ji et al., 2005] In real data analysis, if we assume independence but don't have it, we may make more type I errors based on my experience. [Dai and Charnigo, 2015] introduced compound hierarchical correlated beta mixture model to model the correlation structure of genes. So in the future, we may consider a bayesian hierarchical beta contamination model with distribution of correlation coefficient as prior to deal with the correlation variations among genes.

Copyright© Ya Qi, 2022.

<https://orcid.org/0000-0001-6656-7008>

Bibliography

- [Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.
- [Allison et al., 2002] Allison, D. B., Gadbury, G. L., Heo, M., Fernández, J. R., Lee, C.-K., Prolla, T. A., and Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis*, 39(1):1–20.
- [Benaglia et al., 2009] Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. (2009). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29.
- [Billingsley, 1968] Billingsley, P. (1968). *Convergence of probability measures*. John Wiley & Sons.
- [Byrd et al., 1995] Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.
- [Böhning et al., 1994] Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., and Lindsay, B. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, 46(2):373–388.
- [Charnigo and Sun, 2004] Charnigo, R. and Sun, J. (2004). Testing homogeneity in a mixture distribution via the l2 distance between competing models. *Journal of the American Statistical Association*, 99(466):488498.

- [Charnigo and Sun, 2010] Charnigo, R. and Sun, J. (2010). Asymptotic relationships between the d-test and likelihood ratio-type tests for homogeneity. *Statistica Sinica*, 20(2):497.
- [Charnigo et al., 2010] Charnigo, R. J., Chesnut, L. W., LoBianco, T., and Kirby, R. S. (2010). Thinking outside the curve, part i: modeling birthweight distribution. *BMC Pregnancy and Childbirth*, 10:37 – 37.
- [Chen et al., 2001] Chen, H., Chen, J., and Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(1):19–29.
- [Chen et al., 2004] Chen, H., Chen, J., and Kalbfleisch, J. D. (2004). Testing for a finite mixture model with two components. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 66(1):95–115.
- [Chen et al., 2010] Chen, J., Li, P., and Fu, Y. (2010). Inference on the order of a normal mixture. *Journal of the American Statistical Association*, 107(499):1096–1105.
- [Dacunha-Castelle and Gassiat, 1999] Dacunha-Castelle, D. and Gassiat, E. (1999). Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes. *The Annals of Statistics*, 27(4):1178 – 1209.
- [Dai and Charnigo, 2008] Dai, H. and Charnigo, R. (2008). Omnibus testing and iteration in microarray data analysis. *Journal of Applied Statistics*, 35(1):31–47.

- [Dai and Charnigo, 2010] Dai, H. and Charnigo, R. (2010). Contaminated normal modeling with application to microarray data analysis. *Canadian Journal of Statistics*, 38:315 – 332.
- [Dai and Charnigo, 2015] Dai, H. and Charnigo, R. (2015). Compound hierarchical correlated beta mixture with an application to cluster mouse transcription factor dna binding data. *Biostatistics*, 16(4):641–654.
- [Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.
- [Drton et al., 2009] Drton, M., Eichler, M., and Richardson, T. S. (2009). Computing maximum likelihood estimates in recursive linear models with correlated errors. *Journal of Machine Learning Research*, 10(81):2329–2348.
- [Drton and Plummer, 2017] Drton, M. and Plummer, M. (2017). A bayesian information criterion for singular models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 9(2):323–380.
- [Fletcher, 2000] Fletcher, R. (2000). *The Theory of Constrained Optimization*, chapter 9, pages 195–228. John Wiley Sons, Ltd.
- [Furman and Lindsay, 1994] Furman, W. D. and Lindsay, B. G. (1994). Testing for the number of components in a mixture of normal distributions using moment estimators. *Computational Statistics & Data Analysis*, 17(5):473–492.
- [Jabbari and Bernardi, 2004] Jabbari, K. and Bernardi, G. (2004). Cytosine methylation and cpg, tpg (cpa) and tpa frequencies. *Gene*, 333:143–149.
- [Ji et al., 2005] Ji, Y., Wu, C., Liu, P., Wang, J., and Coombes, K. R. (2005). Applications of beta-mixture models in bioinformatics. *Bioinformatics*, 21(9):2118–2122.

- [Karlis and Xekalaki, 2003] Karlis, D. and Xekalaki, E. (2003). Choosing initial values for the em algorithm for finite mixtures. *Computational Statistics & Data Analysis*, 41(3-4):577–590.
- [Keribin, 2000] Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics*, 62(1):49–66.
- [Lahiri, 2001] Lahiri, P. (2001). Model selection. IMS.
- [Laird, 1978] Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811.
- [Li et al., 2009] Li, P., Chen, J., and Marriott, P. (2009). Non-finite Fisher information and homogeneity: an EM approach. *Biometrika*, 96(2):411–426.
- [McLachlan, 1994] McLachlan, Geoffrey, D. P. (1994). *Finite mixture models*. John Wiley Sons, first edition.
- [Naumova et al., 2021] Naumova, O. Y., Lipschutz, R., Rychkov, S. Y., Zhukova, O. V., and Grigorenko, E. L. (2021). Dna methylation alterations in blood cells of toddlers with down syndrome. *Genes*, 12(8).
- [Ospina and Ferrari, 2012] Ospina, R. and Ferrari, S. L. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, 56(6):1609–1623.
- [Qi, 2016] Qi, M. (2016). Development in normal mixture and mixture of experts modeling.
- [R Core Team,] R Core Team. *R: constrOptim discription*. R Foundation for Statistical Computing, Vienna, Austria.

- [Roeder, 1994] Roeder, K. (1994). A graphical technique for determining the number of components in a mixture of normals. *Journal of the American Statistical Association*, 89(426):487–495.
- [Schwarz, 1978] Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.
- [Watanabe, 2009] Watanabe, S. (2009). *Algebraic Geometry and Statistical Learning Theory*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press.
- [Wieczorek and Hawala, 2011] Wieczorek, J. and Hawala, S. (2011). A bayesian zero-one inflated beta model for estimating poverty in us counties. In *Proceedings of the American Statistical Association, Section on Survey Research Methods, Alexandria, VA: American Statistical Association*.
- [Young et al., 2022] Young, D. S., Roemmele, E. S., and Yeh, P. (2022). Zero-inflated modeling part i: Traditional zero-inflated count regression models, their applications, and computational tools. *Wiley Interdisciplinary Reviews: Computational Statistics*, 14(1):e1541.

Vita

Ya Qi

EDUCATION

PHD in Statistics, University of Kentucky, Lexington, KY Expected Aug 2022

M.S. in Statistics, University of Kentucky, Lexington, KY May 2015

B.S. in Biotechnology, East China Normal University, Shanghai, China May 2011

PROFESSIONAL EXPERIENCE

Senior Statistical Consultant and Analyst — Applied Statistical Lab, University of Kentucky, 2015.8 — 2020.1

Teaching Assistant — University of Kentucky, 2013.8 — 2015.7

PUBLICATIONS AND PRESENTATIONS

- Large-Scale Hypothesis Testing with a Three-Component Beta Contamination Model by Ya Qi and Richard Charnigo. JSM 2017.
- Neonatal Outcomes in Buprenorphine/Naloxone versus Buprenorphine Monotherapy for Medication Assisted Therapy of Opioid Use Disorder. Quinetta B. Johnson, John O'Brien, Greg Hawk, Ya Qi, Agatha Critchfield. American College of Obstetricians and Gynecologists 2019.
- Improvement in CREOG Scores Through Focused Review Sessions by Ian Cook, Miriam Marcum, Yan Xu, Ya Qi, Kristen McQuerry, Christopher DeSimone. CREOG & APGO Meeting 2018.
- Neonatal head circumference in opioid-exposed neonates with and without Neonatal Opioid Withdrawal Syndrome (NOWS) Quinetta B. Johnson, John O'Brien,

Greg Hawk, Ya Qi, Agatha Critchfield. Society of Maternal Fetal Medicine 2018.

- Relationship between buprenorphine dose, timing of buprenorphine dosing to delivery and neonatal opioid withdrawal syndrome. Aarthi Srinivasan, Ya Qi. Society of Maternal Fetal Medicine 2018.
- Ultrasound-indicated cerclage placement for early preterm birth prevention in women without a prior preterm birth. Aarthi Srinivasan, Ya Qi. Society of Maternal Fetal Medicine 2018.
- Can a Low-Fidelity Surgical Model Simulating Loss of Vessel Control During Uterine Artery Ligation Induce Stress Among Gynecologic Residents? Rone, Bryan; Ollendorff, Arthur; Qi, Ya. Obstetrics & Gynecology 2016.
- Comparative analysis of human protein-coding and noncoding RNAs between brain and 10 mixed cell lines by RNA-Seq. Chen G, Yin K, Shi L, Fang Y, Qi Y, Li P, Luo J, He B, Liu M, Shi T. PLoS One, 2011.
- Inhibition of breast cancer metastases by a novel inhibitor of TGF β receptor 1. Y Fang, Y Chen, L Yu, C Zheng, Y Qi, Z Li, Z Yang... - Journal of the National Cancer Institute, 2012.