2014

NORMAL MIXTURE AND CONTAMINATED MODEL WITH NUISANCE PARAMETER AND APPLICATIONS

Qian Fan
University of Kentucky, qfa222@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.

Qian Fan, Student

Dr. Richard J. Charnigo, Jr, Major Professor

Dr. Constance L. Wood, Director of Graduate Studies

NORMAL MIXTURE AND CONTAMINATED MODEL WITH NUISANCE PARAMETER AND APPLICATIONS

ABSTRACT OF 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
Qian Fan
Lexington, Kentucky

Director: Dr. Richard Charnigo, Professor of Statistics
Lexington, Kentucky 2014

ABSTRACT OF DISSERTATION

NORMAL MIXTURE AND CONTAMINATED MODEL WITH NUISANCE PARAMETER AND APPLICATIONS

This paper intend to find the proper hypothesis and test statistic for testing existence of bilaterally contamination when there exists nuisance parameter. The test statistic is based on method of moments estimators. Union-Intersection test is used for testing if the distribution of population can be implemented by a bilaterally contaminated normal model with unknown variance. This paper also developed a hierarchical normal mixture model (HNM) and applied it to birth weight data. EM algorithm is employed for parameter estimation and a singular Bayesian information criterion (sBIC) is applied to choose the number components. We also proposed a singular flexible information criterion which in addition involves a data-driven penalty.

KEYWORDS: bilaterally contaminated normal model, Union-Intersection test, hierarchical normal mixture model, singular Bayesian information criterion, singular flexible information criterion

Author's signature: Qian Fan

Date: December 16, 2014

NORMAL MIXTURE AND CONTAMINATED MODEL WITH NUISANCE PARAMETER AND APPLICATIONS

By
Qian Fan

Director of Dissertation: Richard Charnigo

Director of Graduate Studies: Constance Wood

Date: December 16, 2014

ACKNOWLEDGMENTS

Before I wrote this dissertation, I had concerns and doubts. I had no experience writing anything lengthy and theoretical like this and I didn't know where to start. Not to mention the language barrier that any non-English speaking, foreign student faces. Of course, the hardest part is the theoretical proof for which there was not much previous work that I could refer to. Then, as I began to draft chapter by chapter, I started to have a much better understanding of mixture modeling. At this point, with my dissertation finished, I've grown a lot of skills. Through this experience, I've learned to think and work independently. I'm able to bridge theory and practice, and think critically and outside the frame. I also developed my verbal and writing skills, and learned to explain things to other people.

First of all, I would like to acknowledge my advisor, Dr. Richard Charnigo, for his time and excellent guidance on my dissertation. He also gave me considerable encouragement over the last several years. I have learned a great amount of skills from him. He has been an excellent advisor, as well as a great friend. He is always very patient when I do not understand a theory or proof. We had weekly meetings and I will always learn something new from each meeting. As an international student, my dissertation inevitably has language problems. Dr. Charnigo will read it word by word and correct my mistakes. I cannot thank him enough on the amount of help and advice he has provided on my dissertation. Without him, I could not have accomplished what has been done.

I am very fortunate to have Dr. Arnold Stromberg, Dr. Ruriko Yoshida, Dr. Simon Bonner, Dr. William Griffith, Dr. David Fardo and Dr. Chang-Guo Zhan serve on

my committee. Dr. Stromberg has been very supportive on my dissertation. He also supported me financially during this process so I could focus on my work. Dr. Yoshida has been like a co-advisor in many ways. She has brought brilliant ideas to my dissertation. I'm grateful for Dr. Bonner who agreed to be my committee member when Dr. Yoshida is not able to attend. I'm also grateful for all the support I received from Dr. Griffith, Dr. Fardo and Dr. Zhan.

I am grateful to Hongying Dai, an assistant professor at University of Missouri Kansas City. Professor Dai gave very important support on supplying the data and bringing brilliant ideas to add to my dissertation. I am also grateful to other faculty members, fellow students and staff from Department of Statistics and Department of Biostatistics. I am thankful for all the knowledge I learned from the faculty and peers both inside and outside class. Their help and challenges have been my motivation to continuously do good work on my research.

Last but not least, I want to thank my caring husband, Bradley Glass. You have been so supportive when I have had to work late or feel upset. Thank you for patiently supporting my academic pursuit and sharing a wonderful life with me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

1.1 Finite Mixture Model

Let $X_1, ..., X_n$ denote a random sample of size n, X_i is a p-dimensional random vector with density function $f(x_i)$. We suppose that the density $f(x_i)$ can be written in the form

$$f(x_i) = \sum_{j=1}^{g} \pi_j f_j(x_i),$$

where the $f_j(x_i)$ are densities and the $\pi_{j's} \in [0,1], j = 1, 2, ..., g < \infty, \sum_{j=1}^{g} \pi_j = 1$. In such a case, we shall say that $X_1, ..., X_n$ have a *finite mixture distribution* and that $f(\cdot)$ is a *finite mixture density function*. The parameters $\pi_j's$ are called the *mixing weights* and $f_j(\cdot)$ are called the *component densities* of the mixture. Often, $f_j(\cdot)$ is taken to have the form $g(\cdot;\theta_j)$, where $g(\cdot;\theta_j)$ is a probability density function governed by a parameter θ_j and $\theta_j \in \Theta$, a parameter space. In the model, the number of components g is considered fixed. But in applications, the smallest possible value of g with which $f(x_i)$ can be recovered is unknown and has to be inferred from the available data. For example, the following three mixture models are equivalent, since the second model has a third component which has weight zero while the last two terms of the third model can be combined.

$$0.5N(0,1) + 0.5N(1,1)$$

$$0.5N(0,1) + 0.5N(1,1) + 0N(2,1)$$

$$0.5N(0,1) + 0.25N(1,1) + 0.25N(1,1)$$

Above $N(\mu, \sigma^2)$ represents the density of the normal distribution with mean μ and variance σ^2.

There are many applications of mixture models, Charnigo's (2010a) paper proposed using a normal mixture model to describe a birth weight distribution in which the number of components is determined from the data using a model selection criterion rather than fixed *a priori*. They found that a 4-component normal mixture model reasonably describes the birth weight distribution for a population of white singleton infants born to heavily smoking mothers. Another example is the semiparametric mixture model which is characterized by a non-parametric mixing distribution Ω and a structural parameter β common to all components. In Charnigo's 2006 paper with Pilla, they created a framework for consistent estimation of Ω, β and the order of the model m. The order of a finite mixture model is the smallest number of components with which the mixture density can be recovered. Or in the notation of Charnigo and Pilla (2006), m is the order of the support set of Ω. Suppose we have

$$0.5N(0,1) + 0.5N(1,1).$$

We can express this in terms of Ω and β by defining Ω to be a discrete distribution that places 0.5 of its mass on 0 and 0.5 of its mass on 1:

$$P(U = 0) = 0.5, P(U = 1) = 0.5$$

where $U \sim \Omega$.

$$0.5N(0,1) + 0.5N(1,1) = \sum_{u \in sup(\Omega)} P(U = u)N(u,1)$$

Here, β is set to 1. They formulated a class of generalized exponential family (GEF) models and established sufficient conditions for the identifiability of finite mixtures formed from a GEF along with sufficient conditions for a nesting structure. They found that consistent estimation of the order m is possible if one restricts the class of mixing distributions and employs an information-theoretic approach. Charnigo and Pilla (2006) also exhibited practical applications of mixture modeling, including the

analysis of sodium-lithium countertransport (SLC) data.

Statistical inference in mixture modeling is difficult. Regularity conditions are violated in the mixture problem which result in a loss of identifiability. For instance, $(1-\pi)N(0,1)+\pi N(\mu,1)=N(0,1)$ for any μ if $\pi=0$ or for any π if $\mu=0$. There is not a unique π or a unique μ through which $(1-\pi)N(0,1)+\pi N(\mu,1)$ reduced to $N(0,1)$. The classical log likelihood ratio test statistic does not maintain usual asymptotic structure for testing homogeneity against a mixture alternative. (Note, homogeneity is defined as an order of 1 in a mixture model.) Hartigan (1985) showed that the LRT goes to infinity in probability if the parameter space Θ is unbounded for a normal mixture model. Under the assumption that Θ is bounded, Ghosh and Sen (1985) pointed out if they reparameterize to set up *Euclidean* parameters in one to one correspondence with the mixing distribution, then the problem becomes one of lack of differentiability of the density with respect to these new parameters, at points in the space of H_0. They developed the asymptotic theory for LRT for testing homogeneity by restricting the two mixing parameters to be separated.

Chen and Chen (2001) removed the separation condition of Ghosh and Sen (1985) and stated the asymptotic null distribution of the LRT is the squared supremum of a truncated Gaussian process on Θ and recommended a bootstrap procedure for testing homogeneity. Chen and Chen, in collaboration with Kalbfleisch (2001), also proposed a $modified$ LRT which retains the power of the LRT and has a clear form of asymptotic properties. The test is based on the following modified log-likelihood function

$$l_n(\pi,\theta_1,\theta_2)=\sum_{i=1}^{n}\log\{(1-\pi)f(X_i,\theta_1)+\pi f(X_i,\theta_2)\}+C\log\{4\pi(1-\pi)\}$$

where $0\le\pi\le 1$, $\theta_1,\theta_2\in\Theta$ with $\theta_1\le\theta_2$, $C>0$ is a constant and is used to control the level of modification. l_n is often called a penalized likelihood function, referring

to the penalty term $C\log\{4\pi(1-\pi)\}$ which is large when π is close to 0 or 1. The penalty effectively eliminates the identifiability problem and yields a test statistic whose asymptotic null distribution is the mixture of χ_1^2 and χ_0^2 with equal weights under some regularity conditions, i.e.

$$0.5\chi_1^2 + 0.5\chi_0^2,$$

where χ_0^2 is a degenerate distribution with all its mass at 0.

Chen, Chen and Kalbfleisch (2004) extended the modified likelihood approach to finite normal mixture models with a common and unknown variance in the mixing components and consider a test of the hypothesis of a homogeneous model versus a mixture on two or more components. They found that the χ_2^2 distribution is a stochastic lower bound to the limiting null distribution of the likelihood ratio statistic, meanwhile, its a stochastic upper bound to the limiting distribution of the modified likelihood ratio statistic. But they suggested that in practice, a data analyst could conservatively take critical values from a χ_3^2 distribution.

Charnigo and Sun (2004) proposed another method for testing homogeneity, namely the D-test. It's based on the L^2 distance between the fitted null and alternative models and has closed-form expressions for mixture components from standard distributions. In other words, the D-test statistic can be expressed as a function of $\hat{\pi}_1, ..., \hat{\pi}_g, \hat{\theta}_1, ..., \hat{\theta}_g, \hat{\theta}_0$, where $\hat{\pi}_1, ..., \hat{\pi}_g, \hat{\theta}_1, ..., \hat{\theta}_g$ are estimates of mixture model parameters under the alternative hypothesis with g components and $\hat{\theta}_0$ is an estimator of the (single) model parameter under the null hypothesis of homogeneity. Charnigo and Sun (2004) showed that the D-test is consistent for a 2-component alternative hypothesis when parameters are estimated by maximum likelihood and demonstrated through simulations that the D-test had power comparable to that of the LRT and MLRT. Charnigo and Sun (2010) later showed that the D-test are asymptotically

equivalent to three likelihood ratio-type tests for homogeneity, under maximum likelihood, Bayesian estimation framework and empirical Bayesian estimation framework respectively. The second equivalence yields a simple limiting null distribution of the D-test statistic, which involves an estimable constant $C^*(\theta_0)$ such that $nd_nC^*(\theta_0)^{-1}$ is distributed as $0.5\chi_0^2 + 0.5\chi_1^2$. It also yields a simple limiting distribution under contiguous local alternatives for the D-test statistic, which reveals that the D-test is asymptotically locally most powerful. The third equivalence also yields a simple limiting null distribution for D-test. Under the indicated empirical Bayesian estimation framework, as $n \to \infty, nd_n = A^*(\alpha_0)\{E_n(\alpha_0) - 2\log[2\alpha_0]\} + o_p(1)$, where α_0 is the initial value for α in an iterative algorithm, $E_n(\alpha_0)$ denotes the EM test statistic based on one set of initial values that includes the initial value $\alpha_0 \in (0, 0.5]$ for α (the weight) of the second mixture component, and

$$A^*(\alpha) := \frac{5}{32\pi^{1/2}\sigma_0} \text{ for } \alpha < 0.5, \quad A^*(0.5) := \frac{35}{256\pi^{1/2}\sigma_0}.$$

Above, α is the smaller of π_1 and p_2 in a two component mixture, while σ_0 can be substituted by $\hat{\sigma}_0$ without disturbing the conclusions of this theorem or the following corollary. When $\alpha_0 = 0.5$ the asymptotic null distribution of $nd_n\{256/35\}\pi^{1/2}\sigma_0$ is $0.5\chi_0^2 + 0.5\chi_1^2$.

Chen and Li (2009) proposed an EM-test to finite normal mixture models and showed that the test provides satisfactory solution to three limitations of normal mixture model: the likelihood function is unbounded, the Fisher information is infinite and unidentifiability. The EM test does not require any constraints on the mean and the variance parameters or compactness of the parameter space. They found in finite normal mixture models, when the mixture components share the same variance, the limiting null distribution of EM test is a simple function of the $0.5\chi_0^2 + 0.5\chi_1^2$ and the χ_1^2 distributions. In general normal mixture model, the limiting null distribution of the EM-test is found to be the χ_2^2. Li and Chen (2010) also designed an EM test for

testing the order of a finite mixture model, which is applicable to $H_0 : m = m_0$ versus $H_1 : m > m_0$ for any general order m_0. The null limiting distribution of the EM test has the form of a mixture of $\chi^2_0, \chi^2_1, ..., \chi^2_{m_0}$ distributions, where χ^2_0 is a point mass at 0.

1.2 A Review of Contaminated Density Model

Consider X an observable random variable or vector. For known $k \in \mathbb{N}$ and compact $\Theta_\theta \subset \mathbb{R}^k$, let $\mathscr{F} := \{f_\theta(x) : \theta \in \Theta_\theta\}$ be a known family of distributions. Assume X has the density

$$f(x|\theta, \gamma) := (1 - \gamma)f(x|\theta_0) + \gamma f(x|\theta),$$

where the $\theta \in \Theta_\theta$ and $\gamma \in [0, 1]$ are unknown but fixed values and $\theta_0 \in \Theta_\theta$ is known. We refer to this density as a *contaminated density model* (Dai and Charnigo, 2008b). The difference between a contaminated density model and a mixture model is that θ_0 would also be unknown in a mixture model. To avoid identifiability problem, a mixed parameter $\nu = \gamma(\theta - \theta_0)$ is introduced and we want to test the null hypothesis $H_0 : \nu = 0$ against the complementary alternative hypothesis $H_a : \nu \neq 0$.

Several approaches have been developed to make inference on contamination models. Lemdani and Pons (1999) studied the asymptotic distribution of the LR statistic to test whether the contamination of a known density by another density of the same parametric family reduces to the known density itself. They found that under the null hypothesis, assuming some regularity conditions hold, the likelihood ratio statistic of the above test converges weakly to the supremum of a squared truncated Gaussian process.

Dai and Charnigo (2008b) studied the asymptotic and finite-sample performance of two different tests for contamination, namely a modified likelihood ratio test and

an empirical D-test. They showed that each test statistic has a limiting chi-square distribution under the null hypothesis, while under the complementary alternative hypothesis it has limiting noncentral chi-square distribution. Dai and Chainigo (2008a) also showed these two tests can be used along with a gene filtration process to investigate whether a collection of p-values has arisen from the $Uniform(0,1)$ distribution or whether the $Uniform(0,1)$ distribution contaminated by another $Beta$ distribution is more appropriate.

1.3 Genetic Applications of Mixture or Contaminated Model

Contaminated density models are important and of practical use. In many cases, the parameter θ_0 can be treated as known if we have knowledge on the subject or we have prior experience. It is well known that assuming independence of gene expression levels across genes, the p-values of continuous exact test statistics from a microarray experiment are distributed as $Uniform(0,1)$ under the omnibus null hypothesis of no genome-wise alterations. An exact test is one in which the actual Type I error probability is equal to the nominal Type I error probability at a finite sample size, not just is the limit as n goes to infinity. When you have an exact test based on a continuous test statistics,

$$P(p-values \leq u|H_0) = u$$

for all $u \in (0,1)$. This explains why p-values is distributed as a $Uniform(0,1)$ under null hypothesis.

Parker and Rothenberg (1988) suggested that the distribution of p-values can be fitted by a mixture of a uniform and one or more beta distributions

$$\gamma_0 B(1,1) + \sum_{i=1}^{\nu} \gamma_i B(\alpha_i, \beta_i)$$

where γ_0 is the probability that a randomly chosen test from the collection of tests is for a gene for which there is no population difference in gene expression, and γ_i is the probability that a randomly chosen test from the collection of tests is for a gene from the ith component distribution for which there is a true population difference in gene expression. ν is the number of contaminating components. $\gamma_0 + \sum_{i=1}^{\nu} \gamma_i = 1$ and γ_0 and $\gamma_i \in [0, 1]$.

Allison et al. (2002) adopted this idea and developed a method for modeling the distribution of p-values from microarray experiments when we suspect that there may exist genome-wise alternations. Interestingly, the number of contaminating components is ambiguous unless constraints are placed on the component parameters. For example,

$$0.5B(1,1) + 0.25B(2,1) + 0.25B(1,2)$$

is indistinguishable from $B(1,1)$.

Dai and Charnigo (2008a) showed how to test $H_0 : \nu = 0$ *versus* $H_1 : \nu = 1$ using a modified likelihood ratio test and a D-test, and they proposed that if a collection of p-values is believed to have arisen from the $Uniform(0,1)$ distribution, that collection can be removed from consideration and attention can be directed to a smaller part of the genome. With fewer genes under consideration, investigators may be able to achieve greater power on hypothesis tests while maintaining a desired Type I error rate. If the $Uniform(0,1)$ distribution is contaminated by another Beta distribution, parameter estimates for the contamination model provide a frame of reference for multiple comparisons. For example, if γ_0 is estimated to be 0.5, then the multiple comparisons adjustment may proceed as if the number of genes under consideration were 0.5 times the number of genes actually present.

Dai and Charnigo (2010) also proposed a new approach to analyzing microarray data, which is to use a contaminated normal model $(1-\gamma)N(0,\sigma^2)+\gamma N(\mu,\sigma^2)$ to describe the distribution of Z statistics or suitably transformed T statistics, where $\gamma \in [0,1], \mu \in \mathbb{R}$ and $\sigma^2 \in (0,\infty)$ are unknown but fixed. One may identify $2(1-\Phi|Z_1|),...,2(1-\Phi|Z_n|)$ with $P_1,...,P_n$ which are p-values from microarray experiment. They suggest researchers to analyze Z statistics using contaminated normal model or convert the Z statistics to p-values to be analyzed using the contaminated beta model with $\nu = 1$. If investigator has T statistics, let $\Phi(\cdot)$ denote the cumulative distribution function for a standard normal random variable, and let $\Psi(\cdot)$ denote the cumulative distribution function for a T random variable on appropriate degrees of freedom. Since we can identify $\Phi^{-1}(\Psi(T_1)),...,\Phi^{-1}(\Psi(T_n))$ with $Z_1,...,Z_n$ we may assume without loss of generality that the investigator has Z statistics. They proved that under the omnibus null hypothesis, the MLRT and D-test statistic for the contaminated normal model has a limiting chi-square distribution with one degree of freedom.

In the simulation study, Dai and Charnigo (2010) suggest that contaminated normal model yields more powerful omnibus tests than the contaminated beta model when there is an asymmetry between overexpression and underexpression, or the ratio of $|\mu|$ to σ in the contaminated normal model is not too large. Moreover, contaminated normal model is preferred if one is more concerned about estimating the proportion of differentially expressed genes and if there is an asymmetry between overexpression and underexpression.

A limitation of the contaminated normal model is it does not detect differential expression when the distribution of Z statistics is non-normal but symmetric.

Kendziorski and Newton (2003) proposed a general empirical Bayes modeling approach which allows for replicate expression profiles in multiple conditions. The hierarchical mixture model accounts for differences among genes in their average expression levels, differential expression for a given gene among cell types, and measurement fluctuations., Two distinct parameterizations are considered: a model based on Gamma distributed measurements and one based on log-normally distributed measurements. They also showed how the posterior odds of differential expression in one version of the model is related to the ratio of the arithmetic mean to the geometric mean of the two sample means.

Chapter 2 Bilaterally Contaminated Normal Model without Nuisance Parameter

2.1 Introduction

First, consider the situation where $X_1, ..., X_n$ are independent and identically distributed random variables such that

$$X_1, ..., X_n \sim (1 - \gamma_1 - \gamma_2)N(0, \sigma^2) + \gamma_1 N(\mu_1, \sigma^2) + \gamma_2 N(\mu_2, \sigma^2) \tag{2.1}$$

where $\gamma_1 \in [0, 1-\gamma_2], \gamma_2 \in [0, 1], \mu_1 \in [0, \infty)$ and $\mu_2 \in (-\infty, 0]$ are unknown but fixed and $\sigma^2 \in (0, \infty)$ is known and hereafter assumed to be 1 without loss of generality. We refer to this as the *Bilaterally contaminated normal model without nuisance parameter* (BCN-NP).

In the context of microarray data analysis, we might take X_i to be the following quotient which represents a Z statistic:

$$X_i = \frac{sample\,mean\,on\,patients\,gene\,i - sample\,mean\,on\,controls\,gene\,i}{\sqrt{\frac{sample\,variance\,on\,patients\,gene\,i}{sample\,size_{\,patients}} + \frac{sample\,variance\,on\,controls\,gene\,i}{sample\,size_{\,controls}}}}$$

If the sample size of patients or controls is small, so that the quotient is more appropriately viewed as a T statistic with degree of freedom ν for some $\nu > 0$ rather than a Z statistic, then X_i might be obtained by applying, in succession, the cdf of the T distribution on ν degrees of freedom followed by the inverse standard normal cdf. That is, if we take a quantity X_i that is distributed T_ν, then the evaluation of this quantity at its cdf $\Psi_\nu(X_i)$ is distributed $Unif(0, 1)$ by the probability inverse transformation. Then apply the inverse standard normal cdf to yield $\Phi^{-1}(\Psi_\nu(X_i)) \sim N(0, 1)$.

If $\gamma_1\mu_1$ or $\gamma_2|\mu_2| = 0$, the BCN-NP model reduces to the *Unilaterally contaminated normal model* (UCN-NP). UCN-NP is like the normal mixture model with one contaminating component mentioned in Chapter 1. Consider the situation where $X_1, ..., X_n$ are independent and identically distributed random variables such that

$$X_1, ..., X_n \sim (1-\gamma)N(0, \sigma^2) + \gamma N(\mu, \sigma^2) \tag{2.2}$$

where $\gamma \in [0, 1], \mu \in \mathbb{R}$, and $\sigma^2 = 1$ without loss of generality.

This model can detect either underexpression or overexpression of genes among patients. In microarray data analysis comparing patients with a medical condition to healthy controls, a positive μ indicates some genes have overexpression and a negative μ indicates some genes have underexpression, while γ is the proportion of differential expressed genes. If a gene comes from $N(0, \sigma^2)$, then this gene is neither overexpressed nor underexpressed among patients. Compared to UCN-NP model, BCN-NP model can detect both underexpression and overexpression simultaneously and show how much of each type of differential expression there is. In the BCN-NP model, μ_1 and μ_2 correspond to overexpression and underexpression, respectively, while γ_1 and γ_2 are the proportions of overexpressed and underexpressed genes. This model is handy when both kinds of gene expression are suspected to exist.

The contaminated beta model (CB) (Allison et al 2002; Dai and Charnigo 2008a) has also been employed in microarray data analysis. The model is

$$P_1, ..., P_n \overset{\text{i.i.d.}}{\sim} (1-\gamma)Unif(0,1) + \gamma Beta(\alpha, \beta) \tag{2.3}$$

where $\gamma \in [0, 1]$, $\alpha \in (0, \infty)$, and $\beta \in (0, \infty)$. The CB model describes p-values rather than test statistics as mentioned in Chapter 1. Since both gene underexpression and overexpression produce small p-values($e.g. P_i = 2[1-\Phi|X_i|]$), the CB model can detect both. However, CB model does not distinguish between them. The CB model has

an advantage over the UCN-NP model in that it detects both differential expression types simultaneously, but it does not have the advantage of BCN-NP model that separates the two kinds of differential expression.

2.2 Omnibus Testing Procedure

To determine if both gene under- and over-expression exist, we are interested in testing the two null hypotheses:
(a). Omnibus test:

$$H_0 : N(0,1) \text{ versus } H_a : \text{UCN-NP or BCN-NP} \tag{2.4}$$

(b). Unilateral test:

$$H_0 : \text{UCN-NP versus } H_a : \text{BCN-NP} \tag{2.5}$$

Above we are assuming $\sigma^2 = 1$ without loss of generality.

In this section we study test (a). Test (b) will be studied in section 2.3.

Since the first six moments of the BCN-NP model will be useful for test (a) and/or (b), we record them here:

$$m_1 = \mu_1\gamma_1 + \mu_2\gamma_2 \tag{2.6}$$

$$m_2 = 1 + \mu_1^2\gamma_1 + \mu_2^2\gamma_2 \tag{2.7}$$

$$m_3 = (\mu_1^3 + 3\mu_1)\gamma_1 + (\mu_2^3 + 3\mu_2)\gamma_2 \tag{2.8}$$

$$m_4 = 3 + (\mu_1^4 + 6\mu_1^2)\gamma_1 + (\mu_2^4 + 6\mu_2^2)\gamma_2 \tag{2.9}$$

$$m_5 = (\mu_1^5 + 10\mu_1^3 + 15\mu_1)\gamma_1 + (\mu_2^5 + 10\mu_2^3 + 15\mu_2)\gamma_2 \tag{2.10}$$

$$m_6 = 15 + (\mu_1^6 + 15\mu_1^4 + 45\mu_1^2)\gamma_1 + (\mu_2^6 + 15\mu_2^4 + 45\mu_2^2)\gamma_2 \tag{2.11}$$

Limiting Null Distribution of the Test Statistic

We consider a method of moments approach for testing the null hypothesis a). If this null hypothesis is true, $\gamma_1\mu_1 = \gamma_2|\mu_2| = 0$ and there's no differential expression and further data analysis is not required(e.g., to estimate proportions of differentially expressed genes or to classify which genes are differentially expressed). We propose using $n\hat{m}_2 = \sum_{i=1}^n X_i^2$ as a statistic for testing null hypothesis a) because m_2 has its minimal value(namely, 1) when the null hypothesis is true. Thus, small $\sum_{i=1}^n X_i^2$ can warrant retention of the null hypothesis (a) while large $\sum_{i=1}^n X_i^2$ can warrant rejection. Under null hypothesis (a), $X_1, X_2, ..., X_n \overset{\text{i.i.d.}}{\sim} N(0,1)$, then the statistic $\sum_{i=1}^n X_i^2 \sim \chi_n^2$.

Theorem 2.2.1. *For the statistic $\sum_{i=1}^n X_i^2$, suppose that the null hypothesis a) is true. Then, for any fixed $\alpha \in (0,1)$ we have*

$$P\left(\sum_{i=1}^n X_i^2 > \chi^2_{n,1-\alpha}\right) = \alpha$$

for all n.

Proof. Immediate from the definition of the quantile function. □

Unbiasedness of the Testing Procedure

Now we show that our test for (a) is unbiased, in that rejection of the null hypothesis is more probable when the null hypothesis is false than when true.

Theorem 2.2.2. *For any fixed $\alpha \in (0,1)$, the statistic $\sum_{i=1}^n X_i^2$ satisfies*

$$\mathbb{P}_{H_a}\left(\sum_{i=1}^n X_i^2 > \chi^2_{n,1-\alpha}\right) \geq \mathbb{P}_{H_0}\left(\sum_{i=1}^n X_i^2 > \chi^2_{n,1-\alpha}\right)$$

or equivalently,

$$\mathbb{P}_{H_a}\left(\sum_{i=1}^{n} X_i^2 \leq \chi^2_{n,1-\alpha}\right) \leq \mathbb{P}_{H_0}\left(\sum_{i=1}^{n} X_i^2 \leq \chi^2_{n,1-\alpha}\right).$$

Fact For any fixed n, the $\chi^2_n(\nu)$ distribution, $\nu > 0$, is stochastically greater than the $\chi^2_n(0)$ distribution, where ν represents the non-centrality parameter (Shaked and Shanthikumar 2007). This fact will be useful for the proof of Theorem 2.2.2 below.

Proof. If H_a (a) is true, then

$$\sum_{i=1}^{n} X_i^2 \sim \sum_{i=1}^{3^n} p_i \chi^2_n(\nu_i),$$

where there exists at least one i such that both p_i and ν_i are non-zero and $p_1, ..., p_{3^n} \geq 0$ with $\sum_{i=1}^{3^n} p_i = 1$.

Example If $n = 2$, then there are nine summands with

$p_1 = (1-\gamma_1-\gamma_2)^2,\ \nu_1 = 0,$

$p_2 = (1-\gamma_1-\gamma_2)\gamma_1,\ \nu_2 = \mu_1^2,$

$p_3 = (1-\gamma_1-\gamma_2)\gamma_2,\ \nu_3 = \mu_2^2,$

$p_4 = (1-\gamma_1-\gamma_2)\gamma_1,\ \nu_4 = \mu_1^2,$

$p_5 = \gamma_1^2,\ \nu_5 = 2\mu_1^2,$

$p_6 = \gamma_1\gamma_2,\ \nu_6 = \mu_1^2 + \mu_2^2,$

$p_7 = (1-\gamma_1-\gamma_2)\gamma_2,\ \nu_7 = \mu_2^2,$

$p_8 = \gamma_1\gamma_2,\ \nu_8 = \mu_1^2 + \mu_2^2,$

$p_9 = \gamma_2^2,\ \nu_9 = 2\mu_2^2.$

Indeed, we have $X_1^2 + X_2^2 | Y_1, Y_2 \sim \chi_1^2(Y_1^2) + \chi_2^2(Y_2^2)$ where Y_i is 0 with probability $1-\gamma_1-\gamma_2$, μ_1 with probability γ_1 and μ_2 with probability γ_2. Thus

$$\begin{aligned} X_1^2 + X_2^2 \sim\ & (1-\gamma_1-\gamma_2)^2[\chi_2^2(0)] + 2(1-\gamma_1-\gamma_2)\gamma_1[\chi_2^2(\mu_1^2)] \\ & + 2(1-\gamma_1-\gamma_2)\gamma_2[\chi_2^2(\mu_2^2)] + 2\gamma_1\gamma_2[\chi_2^2(\mu_1^2+\mu_2^2)] \\ & + \gamma_1^2[\chi_2^2(2\mu_1^2)] + \gamma_2^2[\chi_2^2(2\mu_2^2)] \end{aligned} \tag{2.12}$$

So, continuing with the proof,

$$\begin{aligned}
\mathbb{P}_{H_a}(\sum_{i=1}^{n} X_i^2 \leq \chi^2_{n,1-\alpha}) &= \sum_{i=1}^{3^n} p_i \mathbb{P}(\chi^2_n(\nu_i) \leq \chi^2_{n,1-\alpha}) \\
&\leq \sum_{i=1}^{3^n} p_i \mathbb{P}(\chi^2_n(0) \leq \chi^2_{n,1-\alpha}) \qquad (2.13) \\
&= \mathbb{P}(\chi^2_n(0) \leq \chi^2_{n,1-\alpha}) \sum_{i=1}^{3^n} p_i \\
&= \mathbb{P}(\chi^2_n(0) \leq \chi^2_{n,1-\alpha}) \\
&= \mathbb{P}_{H_0}(\sum_{i=1}^{n} X_i^2 \leq \chi^2_{n,1-\alpha}),
\end{aligned}$$

where step (2.13) uses the fact about stochastic ordering. □

Consistency of the Testing Procedure

If null hypothesis (a) is false, then either $\gamma_1\mu_1 \neq 0$ or $\gamma_2|\mu_2| \neq 0$ or both. Under alternative hypothesis (a), $n\,\hat{m}_2 = \sum_{i=1}^{n} X_i^2$ follows a mixture of χ^2_n central and non-central distributions.

Theorem 2.2.3. *For the statistic $\sum_{i=1}^{n} X_i^2$, suppose the alternative hypothesis (a) in the omnibus test is true. Then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{P} 1 + \mu_1^2\gamma_1 + \mu_2^2\gamma_2 > 1. \qquad (2.14)$$

Therefore, for any fixed $\alpha \in (0,1)$ we have

$$\lim_{n\to\infty} P\left(\sum_{i=1}^{n} X_i^2 > \chi^2_{n,1-\alpha}\right) = 1.$$

Proof. Statement (2.14) is an immediate consequence of the weak law of large numbers.

On the other hand, the Central Limit Theorem yields

$$\frac{1}{n}\chi^2_{n,1-\alpha} \longrightarrow E[\chi^2_1] = 1 \; as \; n \to \infty.$$

As such,

$$\lim_{n\to\infty} P_{H_a}(\sum X_i^2 > \chi^2_{n,1-\alpha}) = \lim_{n\to\infty} P_{H_a}\left(\frac{1}{n}\sum X_i^2 > \frac{1}{n}\chi^2_{n,1-\alpha}\right) = 1 \; as \; n \to \infty.$$

□

2.3 Unilateral Testing Procedure

The null hypothesis (b) can be tested using a test statistic whose numerator approximates $M := m_2^2 - 2m_2 + 1 + 3m_1^2 - m_3 m_1 = \gamma_1\gamma_2\mu_1|\mu_2|(\mu_1 - \mu_2)^2$. Under null hypothesis (b), either $\gamma_1\mu_1 = 0$ or $\gamma_2|\mu_2| = 0$ and thus $M = 0$. If the null hypothesis (b) is false, M is strictly positive, since $\gamma_1, \gamma_2 \in (0,1), \mu_1 \in (0,\infty), \mu_2 \in (-\infty, 0)$, so we reject when we get a large test statistic. Define

$$\boldsymbol{m} := (m_1, m_2, ..., m_6)^T, \hat{\boldsymbol{m}} := (\hat{m}_1, \hat{m}_2, ..., \hat{m}_6)^T, \mathbf{0} := (0, 0, ..., 0)^T,$$

$$\mathbf{V(m)} = \begin{pmatrix} m_2 - m_1^2 & m_3 - m_1 m_2 & m_4 - m_1 m_3 \\ m_3 - m_1 m_2 & m_4 - m_2^2 & m_5 - m_2 m_3 \\ m_4 - m_1 m_3 & m_5 - m_2 m_3 & m_6 - m_3^2 \end{pmatrix}. \tag{2.15}$$

Let $g(\mathbf{m}) = m_2^2 - 2m_2 + 1 + 3m_1^2 - m_3 m_1$ and $\mathbf{h(m)} = \frac{\partial g(\mathbf{m})}{\partial \mathbf{m}} = (-m_3 + 6m_1, 2m_2 - 2, -m_1)^T$.

By the multivariate Central Limit Theorem (Ferguson, 1996, p.26), as $n \to \infty$

$$\sqrt{n}(\widehat{\boldsymbol{m}} - \boldsymbol{m}) \xrightarrow{L} N(\mathbf{0}, \boldsymbol{V(m)}). \tag{2.16}$$

By the multivariate Cramer Theorem (Ferguson, 1996, p.45),

$$\sqrt{n}(g(\widehat{\boldsymbol{m}}) - g(\boldsymbol{m})) \xrightarrow{L} N(\mathbf{0}, \boldsymbol{h(m)^T V(m)\, h(m)}). \tag{2.17}$$

Moreover, both $\boldsymbol{h(m)}$ and $\boldsymbol{V(m)}$ are continuous functions of the first six moments. So by Continuous Mapping Theorem and Slutsky's Theorem (Ferguson, 1996)

$$\boldsymbol{h}(\widehat{\boldsymbol{m}})^{\boldsymbol{T}}\, \boldsymbol{V}(\widehat{\boldsymbol{m}})\, \boldsymbol{h}(\widehat{\boldsymbol{m}}) \xrightarrow{P} \boldsymbol{h(m)^T V(m)\, h(m)} \; as \; n \to \infty, \tag{2.18}$$

Provided $\boldsymbol{h(m)^T V(m)\, h(m)} \neq 0$, which is true if null hypothesis (a) is false, this yields

$$\frac{g(\widehat{\boldsymbol{m}}) - g(\boldsymbol{m})}{\sqrt{\boldsymbol{h}(\widehat{\boldsymbol{m}})^{\boldsymbol{T}}\, \boldsymbol{V}(\widehat{\boldsymbol{m}})\, \boldsymbol{h}(\widehat{\boldsymbol{m}})/n}} \xrightarrow{L} N(0,1) \; as \; n \to \infty. \tag{2.19}$$

Now define $Z_n := g(\widehat{\boldsymbol{m}})/\sqrt{\boldsymbol{h}(\widehat{\boldsymbol{m}})^{\boldsymbol{T}}\, \boldsymbol{V}(\widehat{\boldsymbol{m}})\, \boldsymbol{h}(\widehat{\boldsymbol{m}})/n}$, and z_u denote the u quantile of the standard normal distribution (e.g., $z_{0.95} = 1.645$), and $\boldsymbol{m_a}$ be the vector of moments implied by $(\gamma_1, \gamma_2, \mu_1, \mu_2)^T = (\gamma_{1,a}, \gamma_{2,a}, \mu_{1,a}, 0)^T$ for fixed positive constants $\gamma_{1,a}, \gamma_{2,a}$ and $\mu_{1,a}$.

We propose to reject the unilateral null hypothesis (b) if $Z_n > z_{1-\alpha}$. Under the unilateral null hypothesis (b), $\boldsymbol{g(m)} = 0$, $Z_n \xrightarrow{L} N(0,1)$. Obviously, $\lim_{n\to\infty} P(Z_n > z_{1-\alpha}) = \alpha$, so this testing procedure is approximately level α.

Under the local alternative sequence $(\gamma_1, \gamma_2, \mu_1, \mu_2) = (\gamma_{1,a}, \gamma_{2,a}, \mu_{1,a}, -\tau n^{-1/2})$ for a fixed positive constant τ, $\boldsymbol{g(m)} > 0$, then by symmetry of normal distribution,

$$\begin{aligned}
\lim_{n\to\infty} P(Z_n > z_{1-\alpha}) &= \lim_{n\to\infty} P\left(Z_n - \frac{g(\boldsymbol{m})}{\sqrt{\boldsymbol{H}(\widehat{\boldsymbol{m}})/n}} > z_{1-\alpha} - \frac{g(\boldsymbol{m})}{\sqrt{\boldsymbol{H}(\widehat{\boldsymbol{m}})/n}}\right) && (2.20)\\
&= \lim_{n\to\infty} P\left(-Z_n + \frac{g(\boldsymbol{m})}{\sqrt{\boldsymbol{H}(\widehat{\boldsymbol{m}})/\boldsymbol{n}}} \leq -z_{1-\alpha} + \frac{g(\boldsymbol{m})}{\sqrt{\boldsymbol{H}(\widehat{\boldsymbol{m}})/\boldsymbol{n}}}\right) && (2.21)\\
&= \lim_{n\to\infty} P\left(-Z_n + \frac{\gamma_{1,a}\,\gamma_{2,a}\,\mu_{1,a}\,|-\frac{\tau}{\sqrt{n}}|(\mu_{1,a}+\frac{\tau}{\sqrt{n}})^2}{\sqrt{\boldsymbol{H}(\widehat{\boldsymbol{m}})/\boldsymbol{n}}}\right.\\
&\left. \leq -z_{1-\alpha} + \frac{\gamma_{1,a}\,\gamma_{2,a}\,\mu_{1,a}\,|-\frac{\tau}{\sqrt{n}}|(\mu_{1,a}+\frac{\tau}{\sqrt{n}})^2}{\sqrt{\boldsymbol{H}(\widehat{\boldsymbol{m}})/\boldsymbol{n}}}\right) && (2.22)\\
&= \lim_{n\to\infty} P\left(-Z_n + \frac{\gamma_{1,a}\,\gamma_{2,a}\,\mu_{1,a}^3\,\tau}{\sqrt{\boldsymbol{H}(\widehat{\boldsymbol{m}})}} \leq -z_{1-\alpha} + \frac{\gamma_{1,a}\,\gamma_{2,a}\,\mu_{1,a}^3\,\tau}{\sqrt{\boldsymbol{H}(\widehat{\boldsymbol{m}})}}\right) && (2.23)\\
&= \Phi\left(-z_{1-\alpha} + \frac{\gamma_{1,a}\,\gamma_{2,a}\,\mu_{1,a}^3\,\tau}{\sqrt{\boldsymbol{H}(\boldsymbol{m_a})}}\right) && (2.24)
\end{aligned}$$

where $\boldsymbol{H(m)} := \boldsymbol{h(m)^T V(m)\, h(m)}$. Thus, the procedure is asymptotically locally unbiased.

This procedure is also consistent under a fixed alternative, since

$$P(Z_n > z_{1-\alpha}) = P\left(Z_n - \frac{g(\boldsymbol{m})}{\sqrt{\boldsymbol{h}(\widehat{\boldsymbol{m}})^{\boldsymbol{T}}\boldsymbol{V}(\widehat{\boldsymbol{m}})\boldsymbol{h}(\widehat{\boldsymbol{m}})/n}} > z_{1-\alpha} - \frac{g(\boldsymbol{m})}{\sqrt{\boldsymbol{h}(\widehat{\boldsymbol{m}})^{\boldsymbol{T}}\boldsymbol{V}(\widehat{\boldsymbol{m}})\boldsymbol{h}(\widehat{\boldsymbol{m}})/n}}\right).$$

Noting that for any negative real number y we have $P(z_{1-\alpha} - \frac{g(\boldsymbol{m})}{\sqrt{\boldsymbol{h}(\widehat{\boldsymbol{m}})^{\boldsymbol{T}}\boldsymbol{V}(\widehat{\boldsymbol{m}})\boldsymbol{h}(\widehat{\boldsymbol{m}})/n}} < y) \to 1$, we find for sufficiently large n,

$$P(Z_n > z_{1-\alpha}) \geq P(Z_n - \frac{g(\boldsymbol{m})}{\sqrt{\boldsymbol{h}(\widehat{\boldsymbol{m}})^{\boldsymbol{T}}\boldsymbol{V}(\widehat{\boldsymbol{m}})\boldsymbol{h}(\widehat{\boldsymbol{m}})/n}} > y) + \frac{1}{y}. \tag{2.25}$$

The right hand side of inequality (2.25) converges to $\int_y^\infty f(t)\,\mathrm{d}t + \frac{1}{y}$ as $n \to \infty$, where $f(\cdot)$ is the pdf of standard normal distribution. We have

$$1 \geq \limsup_{n\to\infty} P(Z_n > z_{1-\alpha}) \geq \liminf_{n\to\infty} P(Z_n > z_{1-\alpha}) \geq \int_y^\infty f(t)\,\mathrm{d}t + \frac{1}{y} \tag{2.26}$$

for any negative real number y. Since the right hand side of inequality (2.26) converges to 1 as $y \to -\infty$, we can conclude that $\lim_{n\to\infty} P(Z_n > z_{1-\alpha})$ exists and equals 1.

Remark (2.24) is useful for approximate sample size calculations. For example, suppose the local alternative sequence is $(0.1, 0.1, 2, -\tau\, n^{-1/2})$, from (2.6)-(2.11) and the definition of $\boldsymbol{g(m)}, \boldsymbol{V(m)}$ and $\boldsymbol{h(m)}$, one can calculate $m_1 = 0.2, m_2 = 1.4, m_3 = 1.4, m_4 = 7, m_5 = 14.2, m_6 = 63.4, \boldsymbol{h(m_a)} = (-0.2, 0.8, -0.2)$ and

$$\boldsymbol{V(m_a)} = \begin{pmatrix} 1.36 & 1.12 & 6.72 \\ 1.12 & 5.04 & 12.24 \\ 6.72 & 12.24 & 61.44 \end{pmatrix}. \tag{2.27}$$

So $\boldsymbol{h(m_a)^T V(m_a) h(m_a)} = 2$.

When $|\mu_2|$ is small, if 80% power is required, the approximate sample size at level 0.05 is

$$\frac{\boldsymbol{h(m_a)^T V(m_a) h(m_a)}(z_{0.8} + z_{0.95})^2}{\gamma_1^2\,\gamma_2^2\,\mu_1^6\,\mu_2^2} = \frac{1932}{\mu_2^2}.$$

Analogues to (2.24) are also useful when local alternative sequences have γ_1, γ_2 or μ_1 tending to 0. For example, when the sequence is $(\tau\, n^{-1/2}, 0.1, 2, -1)$, the $\gamma_{1,a}\,\gamma_{2,a}\,\mu_{1,a}^3\,\tau$

in (2.24) becomes $\gamma_{2,a}\,\mu_{1,a}\,|\mu_{2,a}|\,(\mu_{1,a}-\mu_{2,a})^2\,\tau$, and $m_1 = -0.1, m_2 = 1.1, m_3 = -0.4, m_4 = 3.7, m_5 = -2.6, m_6 = 21.1, \boldsymbol{h(m_a)} = (-0.2, 0.2, 0.1)$ and

$$\boldsymbol{V(m_a)} = \begin{pmatrix} 1.09 & -0.29 & 3.66 \\ -0.29 & 2.49 & -2.16 \\ 3.66 & -2.16 & 20.94 \end{pmatrix}. \tag{2.28}$$

So $\boldsymbol{h(m_a)^T V(m_a) h(m_a)} = 0.143$.

The approximate sample size is

$$\frac{\boldsymbol{h(m_a)^T V(m_a) h(m_a)}(z_{0.8}+z_{0.95})^2}{\gamma_1^2\,\gamma_2^2\,\mu_1^2\,\mu_2^2\,(\mu_1-\mu_2)^4} = \frac{0.27}{\gamma_1^2}$$

for 80% power.

Similarly, when the sequence is $(0.1, \tau\,n^{-1/2}, 2, -1)$ or $(0.1, 0.1, \tau\,n^{-1/2}, -1)$, the approximate sample sizes are

$$\frac{\boldsymbol{h(m_a)^T V(m_a) h(m_a)}(z_{0.8}+z_{0.95})^2}{\gamma_1^2\,\gamma_2^2\,\mu_1^2\,\mu_2^2\,(\mu_1-\mu_2)^4} = \frac{3.81}{\gamma_2^2}$$

and

$$\frac{\boldsymbol{h(m_a)^T V(m_a) h(m_a)}(z_{0.8}+z_{0.95})^2}{\gamma_1^2\,\gamma_2^2\,\mu_1^2\,\mu_2^6} = \frac{8841}{\mu_1^2}.$$

Remark One can do the test without the restriction that μ_1 is nonnegative and μ_2 is nonpositive. Though the BCN-NP model assumes μ_1 is nonnegative and μ_2 is nonpositive and the proofs above are based on this restriction, this restriction can be lifted by making the rejection region two-sided. Explicitly, if we let μ_1 and μ_2 be arbitrary, we will reject the null hypothesis (b) when we observe $|Z_n|$ greater than $z_{1-\alpha/2}$. In such case, we give the BCN-NP model another name, the *doubly contaminated normal* (DCN) model.

2.4 Parameter Estimation Using Method of Moments

To estimate parameters for the BCN-NP model, we employ a reparameterization. Since we assume μ_1 to be nonnegative and μ_2 to be nonpositive, we can rewrite μ_1

as μ and μ_2 as $-p\mu$, where $\mu \in [0, \infty)$ and $p \in [0, \infty)$. The model becomes

$$X_1, ..., X_n \sim (1 - \gamma_1 - \gamma_2)N(0, 1) + \gamma_1 N(\mu, 1) + \gamma_2 N(-p\mu, 1). \tag{2.29}$$

We will use the method of moments to estimate the parameters.

Define $\widehat{\mu}, \widehat{p}, \widehat{\gamma_1}$ and $\widehat{\gamma_2}$ to satisfy the following equations:

$$\widehat{g_1} := \widehat{m_1} = \widehat{\mu}(\widehat{\gamma_1} - \widehat{p}\widehat{\gamma_2}) \tag{2.30}$$

$$\widehat{g_2} := \widehat{m_2} - 1 = \widehat{\mu}^2(\widehat{\gamma_1} + \widehat{p}^2\widehat{\gamma_2}) \tag{2.31}$$

$$\widehat{g_3} := \widehat{m_3} - 3\widehat{m_1} = \widehat{\mu}^3(\widehat{\gamma_1} - \widehat{p}^3\widehat{\gamma_2}) \tag{2.32}$$

$$\widehat{g_4} := \widehat{m_4} - 6\widehat{m_2} + 3 = \widehat{\mu}^4(\widehat{\gamma_1} + \widehat{p}^4\widehat{\gamma_2}) \tag{2.33}$$

Assuming $\hat{\mu} \neq 0$, let $\widehat{q_i} := \widehat{g_i}/\widehat{\mu}$. Then (2.30) through (2.33) yield the linear systems

$$\begin{bmatrix} \widehat{q_1} \\ \widehat{q_2} \end{bmatrix} = \begin{bmatrix} 1 & -\widehat{p} \\ 1 & \widehat{p}^2 \end{bmatrix} \begin{bmatrix} \widehat{\gamma_1} \\ \widehat{\gamma_2} \end{bmatrix} \tag{2.34}$$

and

$$\begin{bmatrix} \widehat{q_3} \\ \widehat{q_4} \end{bmatrix} = \begin{bmatrix} 1 & -\widehat{p}^3 \\ 1 & \widehat{p}^4 \end{bmatrix} \begin{bmatrix} \widehat{\gamma_1} \\ \widehat{\gamma_2} \end{bmatrix}. \tag{2.35}$$

Assuming $\hat{p} \neq 0$, the two linear systems yield solutions for $\widehat{\gamma_1}$ and $\widehat{\gamma_2}$ given $\widehat{p}$ and $\widehat{\mu}$,

$$\begin{bmatrix} \widehat{\gamma_1} \\ \widehat{\gamma_2} \end{bmatrix} = \begin{bmatrix} 1 & -\widehat{p} \\ 1 & \widehat{p}^2 \end{bmatrix}^{-1} \begin{bmatrix} \widehat{q_1} \\ \widehat{q_2} \end{bmatrix} = \begin{bmatrix} (\widehat{q_2} + \widehat{p}\widehat{q_1})/(1 + \widehat{p}) \\ (\widehat{q_2} - \widehat{q_1})/(\widehat{p} + \widehat{p}^2) \end{bmatrix} \tag{2.36}$$

and

$$\begin{bmatrix} \widehat{\gamma_1} \\ \widehat{\gamma_2} \end{bmatrix} = \begin{bmatrix} 1 & -\widehat{p}^3 \\ 1 & \widehat{p}^4 \end{bmatrix}^{-1} \begin{bmatrix} \widehat{q_3} \\ \widehat{q_4} \end{bmatrix} = \begin{bmatrix} (\widehat{q_4} + \widehat{p}\widehat{q_3})/(1 + \widehat{p}) \\ (\widehat{q_4} - \widehat{q_3})/(\widehat{p}^3 + \widehat{p}^4) \end{bmatrix}. \tag{2.37}$$

Based on (2.36) and (2.37), we wish to equate

$$\widehat{q_2} + \widehat{p}\widehat{q_1} = \widehat{q_4} + \widehat{p}\widehat{q_3}, \tag{2.38}$$

and

$$\widehat{p}^2(\widehat{q_2} - \widehat{q_1}) = \widehat{q_4} - \widehat{q_3}. \tag{2.39}$$

Solving for $\widehat{p}$ given $\widehat{\mu}$ from (2.39) yields

$$\widehat{p} = \sqrt{\frac{\widehat{q_4} - \widehat{q_3}}{\widehat{q_2} - \widehat{q_1}}}. \tag{2.40}$$

To find $\widehat{\mu}$, substitute (2.40) into and multiply (2.39) by $\widehat{\mu}^{10}$, we obtain

$$(\widehat{g_2}\widehat{\mu}^2 - \widehat{g_4})^2(\widehat{g_2} - \widehat{g_1}\widehat{\mu}) = (\widehat{g_3} - \widehat{g_1}\widehat{\mu}^2)^2(\widehat{g_4} - \widehat{g_3}\widehat{\mu}). \tag{2.41}$$

Finally, this yields a quintic equation $\sum_{j=0}^{5} \widehat{h_j}\widehat{\mu}^j = 0$ where

$$\widehat{h_0} := \widehat{g_3}^2\widehat{g_4} - \widehat{g_4}^2\widehat{g_2}, \tag{2.42}$$

$$\widehat{h_1} := \widehat{g_4}^2\widehat{g_1} - \widehat{g_3}^3 \tag{2.43}$$

$$\widehat{h_2} := 2\widehat{g_4}\widehat{g_2^2} - 2\widehat{g_1}\widehat{g_3}\widehat{g_4} \tag{2.44}$$

$$\widehat{h_3} := 2\widehat{g_1}\widehat{g_3}^2 - 2\widehat{g_1}\widehat{g_2}\widehat{g_4} \tag{2.45}$$

$$\widehat{h_4} := \widehat{g_1}^2\widehat{g_4} - \widehat{g_2}^3 \tag{2.46}$$

$$\widehat{h_5} := \widehat{g_1}\widehat{g_2}^2 - \widehat{g_1}^2\widehat{g_3}. \tag{2.47}$$

Then we can solve the quintic equation for $\hat{\mu}$ and evaluate $\widehat{p}$ and $\widehat{\gamma_1}, \widehat{\gamma_2}$.

However, one issue about this approach is the number of real roots of quintic equation. If there is more than one real root, we will look at whether $\hat{\mu}, \hat{p}, \hat{\gamma_1}, \hat{\gamma_2}$ belong to their respective parameter spaces and reproduce the sample moments. First $\hat{\mu}, \hat{p}, \hat{\gamma_1}, \hat{\gamma_2}$ should be real numbers. Then, $\hat{\mu}, \hat{p}$ should be greater than zero and $\hat{\gamma_1}, \hat{\gamma_2}, 1 - \hat{\gamma_1} - \hat{\gamma_2}$ should be between 0 and 1. Moreover, equations (2.30) through (2.33) must be satisfied. These conditions typically yield at most one viable solution of the quintic equation.

We would also like to construct confidence intervals for $\mu, p, \gamma_1, \gamma_2$. If we can derive

a limiting distribution for

$$\sqrt{n}\left(\begin{bmatrix}\widehat{\mu}\\ \widehat{p}\\ \widehat{\gamma_1}\\ \widehat{\gamma_2}\end{bmatrix} - \begin{bmatrix}\mu\\ p\\ \gamma_1\\ \gamma_2\end{bmatrix}\right) \tag{2.48}$$

by Cramer Theorem, we can find $100(1-\alpha)\%$ confidence intervals for these parameters.

Let $\boldsymbol{g(m)} = (\mu, p, \gamma_1, \gamma_2)^T$, and

$$\boldsymbol{A(m)} := \frac{\partial \boldsymbol{g(m)}}{\partial \boldsymbol{m}} = \begin{bmatrix} \frac{\partial \mu}{\partial m_1} & \frac{\partial \mu}{\partial m_2} & \frac{\partial \mu}{\partial m_3} & \frac{\partial \mu}{\partial m_4} \\ \frac{\partial p}{\partial m_1} & \frac{\partial p}{\partial m_2} & \frac{\partial p}{\partial m_3} & \frac{\partial p}{\partial m_4} \\ \frac{\partial \gamma_1}{\partial m_1} & \frac{\partial \gamma_1}{\partial m_2} & \frac{\partial \gamma_1}{\partial m_3} & \frac{\partial \gamma_1}{\partial m_4} \\ \frac{\partial \gamma_2}{\partial m_1} & \frac{\partial \gamma_2}{\partial m_2} & \frac{\partial \gamma_2}{\partial m_3} & \frac{\partial \gamma_2}{\partial m_4} \end{bmatrix}. \tag{2.49}$$

If we can find $\boldsymbol{A(m)}$, we can apply multivariate Cramer Theorem.

Define

$$\boldsymbol{B(\theta)} := \begin{bmatrix} \frac{\partial m_1}{\partial \mu} & \frac{\partial m_1}{\partial p} & \frac{\partial m_1}{\partial \gamma_1} & \frac{\partial m_1}{\partial \gamma_2} \\ \frac{\partial m_2}{\partial \mu} & \frac{\partial m_2}{\partial p} & \frac{\partial m_2}{\partial \gamma_1} & \frac{\partial m_2}{\partial \gamma_2} \\ \frac{\partial m_3}{\partial \mu} & \frac{\partial m_3}{\partial p} & \frac{\partial m_3}{\partial \gamma_1} & \frac{\partial m_3}{\partial \gamma_2} \\ \frac{\partial m_4}{\partial \mu} & \frac{\partial m_4}{\partial p} & \frac{\partial m_4}{\partial \gamma_1} & \frac{\partial m_4}{\partial \gamma_2} \end{bmatrix}. \tag{2.50}$$

Define $\boldsymbol{\theta_0} := (\mu_0, p_0, \gamma_{1,0}, \gamma_{2,0})^T$, and $\boldsymbol{m_0} := (m_{1,0}, m_{2,0}, m_{3,0}, m_{4,0})^T$ to be the "true" parameter and moments respectively. By the Inverse Function Theorem, for $\boldsymbol{\theta} = (\mu, p, \gamma_1, \gamma_2)^T$, $\boldsymbol{m} = (m_1, m_2, m_3, m_4)^T$, if $\boldsymbol{B(\theta_0)}$ has non-zero determinant, then there exists a neighborhood D about $\boldsymbol{\theta_0}$ such that the map $\boldsymbol{h(\theta)} = \boldsymbol{m}$ is invertible on D. The inverse map $\mathbf{g(m)}$=$\boldsymbol{\theta}$ is differentiable. So $\boldsymbol{A(m_0)} = \boldsymbol{B(\theta_0)}^{-1}$, where $\boldsymbol{m_0} = \boldsymbol{h(\theta_0)}$. Then by the multivariate Cramer Theorem, we have

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}) \xrightarrow{L} N(\mathbf{0}, \boldsymbol{A(m_0)V A(m_0)^T}) = N(\mathbf{0}, \boldsymbol{B(\theta_0)^{-1}V(B(\theta_0)^{-1})^T})$$

where $\boldsymbol{V}$ is from section 2.3. The estimated elements of $\boldsymbol{B(\theta_0)^{-1}V(B(\theta_0)^{-1})^T}$, can be used to obtain $100(1-\alpha)\%$ confidence intervals. More specifically, let $(\widehat{\rho_1}, \widehat{\rho_2}, \widehat{\rho_3}, \widehat{\rho_4})^T$ denote the estimated diagonal elements of $\boldsymbol{B(\theta_0)^{-1}V(B(\theta_0)^{-1})^T}$, then the 95% confidence intervals for $\mu, p, \gamma_1, \gamma_2$ are $\widehat{\mu} \pm Z_{0.975}\sqrt{\widehat{\rho_1}/n}$, $\widehat{p} \pm Z_{0.975}\sqrt{\widehat{\rho_2}/n}$, $\widehat{\gamma_1} \pm Z_{0.975}\sqrt{\widehat{\rho_3}/n}$ and $\widehat{\gamma_2} \pm Z_{0.975}\sqrt{\widehat{\rho_4}/n}$.

2.5 Simulation Study

To see if there is one and only one viable solution to the quintic equation in section 2.4, some simulation study is needed. Let $\mu \in \{1, 2, 4\}, p \in \{0.5, 1, 2\}, \gamma_1, \gamma_2 \in \{0.05, 0.1, 0.2\}$ and sample size $n \in \{1000, 5000, 10000\}$. At each of the 243 combinations of $\mu, p, \gamma_1, \gamma_2, n$, 1000 data sets were generated. More specifically, to create each data set, we randomly generate $U_1, ..., U_n$ from $Unif(0, 1)$. If $U_j < \gamma_1$, then X_j is simulated from $N(\mu, 1)$, if $U_j > 1 - \gamma_2$, then X_j is simulated from $N(-p\mu, 1)$ and if $\gamma_1 < U_j < 1 - \gamma_2$, then X_j is simulated from $N(0, 1)$. Then viability of solutions are tested as described in section 2.4. Partial results indicating the numbers of viable solutions among 1000 data sets from various scenarios are given below:

Table 2.1: Numbers of samples that produce 0, 1, 2, 3, 4, 5 viable roots

μ	p	γ_1	γ_2	n	0	1	2	3	4	5
1	0.5	0.05	0.05	1000	747	253	0	0	0	0
1	0.5	0.05	0.05	5000	547	452	1	0	0	0
1	0.5	0.05	0.05	10000	446	553	1	0	0	0
2	0.5	0.05	0.05	1000	251	749	0	0	0	0
2	0.5	0.05	0.05	5000	22	978	0	0	0	0
2	0.5	0.05	0.05	10000	1	999	0	0	0	0
4	0.5	0.05	0.05	1000	0	1000	0	0	0	0
4	0.5	0.05	0.05	5000	0	1000	0	0	0	0
4	0.5	0.05	0.05	10000	0	1000	0	0	0	0
		.....					.....			

Table 2.2: Numbers of samples that have 0, 1, 2, 3, 4, 5 roots yielding non-real parameter estimates

μ	p	γ_1	γ_2	n	0	1	2	3	4	5
1	0.5	0.05	0.05	1000	243	0	445	0	312	0
1	0.5	0.05	0.05	5000	422	0	357	0	221	0
1	0.5	0.05	0.05	10000	515	0	350	0	135	0
2	0.5	0.05	0.05	1000	790	0	205	0	5	0
2	0.5	0.05	0.05	5000	994	0	6	0	0	0
2	0.5	0.05	0.05	10000	1000	0	0	0	0	0
4	0.5	0.05	0.05	1000	1000	0	0	0	0	0
4	0.5	0.05	0.05	5000	1000	0	0	0	0	0
4	0.5	0.05	0.05	10000	1000	0	0	0	0	0
		.....					.....			

Table 2.3: Numbers of samples that have 0, 1, 2, 3, 4, 5 roots failing to satisfy equations (2.30) through (2.33)

μ	p	γ_1	γ_2	n	0	1	2	3	4	5
1	0.5	0.05	0.05	1000	0	0	0	819	0	181
1	0.5	0.05	0.05	5000	0	0	2	818	0	180
1	0.5	0.05	0.05	10000	0	1	4	787	0	208
2	0.5	0.05	0.05	1000	0	0	1	815	0	184
2	0.5	0.05	0.05	5000	0	0	0	986	0	14
2	0.5	0.05	0.05	10000	0	0	0	999	0	1
4	0.5	0.05	0.05	1000	0	0	0	1000	0	0
4	0.5	0.05	0.05	5000	0	0	0	1000	0	0
4	0.5	0.05	0.05	10000	0	0	0	1000	0	0
		.....					.....			

Table 2.4: Numbers of samples that have 0, 1, 2, 3, 4, 5 roots yielding real parameter estimates ourside of their respective parameter spaces

μ	p	γ_1	γ_2	n	0	1	2	3	4	5
1	0.5	0.05	0.05	1000	0	0	91	179	218	512
1	0.5	0.05	0.05	5000	0	0	185	283	234	298
1	0.5	0.05	0.05	10000	0	0	236	369	183	212
2	0.5	0.05	0.05	1000	0	0	212	635	124	29
2	0.5	0.05	0.05	5000	0	0	13	975	12	0
2	0.5	0.05	0.05	10000	0	0	0	1000	0	0
4	0.5	0.05	0.05	1000	0	0	0	1000	0	0
4	0.5	0.05	0.05	5000	0	0	0	1000	0	0
4	0.5	0.05	0.05	10000	0	0	0	1000	0	0
		.....					.....			

Among 243 combinations of parameters and sample sizes, 184 had more than 950 out of 1000 samples that produced one and only one viable solution, while 199 had more than 850 out of 1000 samples that produced one and only one viable solution. So when the data were generated from BCN-NP model, there was often a unique root to the quintic equation yielding real parameter estimates that belonged to their respective parameter spaces and that satisfied equations (2.30) through (2.33).

Test statistics were also calculated for each sample. For omnibus test, the test statistic $n\,\widehat{m_2}$ was compared to $\chi^2_{n,0.95}$; we rejected the null hypothesis if the observed test statistic was larger than $\chi^2_{n,0.95}$. For unilateral test, the test statistic Z_n was compared to $z_{0.95}$; we rejected the null hypothesis if the observed Z_n was greater than $z_{0.95}$.

Among the 243 combinations, there were 240 where more than 800 out of 1000 samples rejected the omnibus null hypothesis and 217 where more than 800 out of 1000 samples rejected the unilateral null hypothesis, when in fact the samples followed a BCN-NP model that could not be reduced to a UCN-NP model or to a single normal distribution. Generally, the omnibus test has more power than the unilateral test.

2.6 Application to Down's Syndrome Microarray Data Set

In Mao's 2005 paper, the microarray data of Down's syndrome patients were analyzed and they are available at http://www.partek.com/Tutorials. There are four patients with Down's syndrome and four healthy people as the controls. Four samples of cerebral tissue were taken from each patient and seven samples from four controls, two from each of the first three and one from the last person. It is known that the Down's syndrome patient has an extra chromosome 21. So we specifically looked at 251 genes available on chromosome 21, in addition to all of the genes on all chromosomes.

We fit a linear mixed model on each of the genes. Let $i = 1, ..., 8, j = 1, 2$ index subjects and samples, the model is

$$y_{ij} = \beta_0 + \beta_1 x_i + \alpha_i + \epsilon_{ij}$$

where y_{ij} denotes the expression level for subject i and sample j; x_i is the indicator for Down's syndrome (1 for yes, 0 for no); β_0 and β_1 are the intercept and coefficient for fixed effect; α_i is a random effect for subject i, and ϵ_{ij} is the error term for subject i and sample j.

Coefficients and variance components were estimated by `lme` function in `R`. The Z test statistics modeled with UCN and BCN models were transformed versions of the T test statistics for $\beta_1 = 0$ at each gene. For all the chromosomes, the fitted UCN model had $(\hat{\gamma}, \hat{\mu}) = (0.49, -1.16)$, while fitted BCN model had $(\hat{\gamma_1}, \hat{\gamma_2}, \hat{\mu_1}, \hat{\mu_2}) = (0.75, 0.25, -0.96, 0.89)$. From Figure 2.1, the distribution is unimodal and moderately skewed to the right. The BCN model may not show a bimodal shape, but it fits the data better than UCN model on the test statistics between [-1, 1]. The unilateral null hypothesis was rejected with Z=21.5.

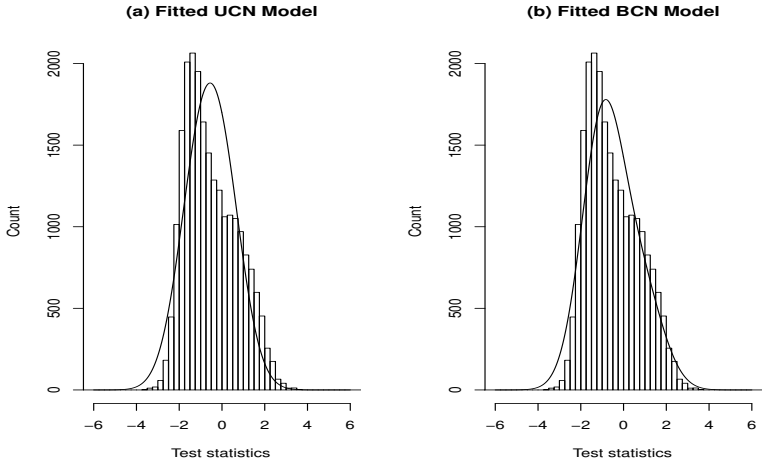

Figure 2.1: The fitted UCN and BCN models on all chromosomes.

However, when we looked at chromosome 21, the results were noticeably different from those obtained overall and at other individual chromosomes. To demonstrate, we compared chromosome 21 with chromosome 10, which we took as representative of the other individual chromosomes. Figures 2.2 and 2.3 are the fitted UCN and BCN models on chromosomes 10 and 21. Apparently, chromosome 21 exhibits abnormal pattern, the empirical distribution of transformed T test statistics was bimodal. The fitted BCN model was obviously better than fitted UCN model at describing the empirical distribution of transformed T test statistics.

For 783 available genes of chromosome 10, the Z test statistic as defined in Section 2.3 was 0.00, so we failed to reject the unilateral null hypothesis. The fitted UCN model had $(\hat{\gamma}, \hat{\mu}) = (0.378, -1.2)$, while fitted BCN model gives $(\hat{\gamma_1}, \hat{\gamma_2}, \hat{\mu_1}, \hat{\mu_2}) = (0.34, 0.66, 0.74, -0.95)$. For 251 available genes of chromosome 21, the Z test statistic is 5.84, so the unilateral null hypothesis was rejected. The fitted UCN model gives $(\hat{\gamma}, \hat{\mu}) = (0.292, 2.41)$, while fitted BCN model gives $(\hat{\gamma_1}, \hat{\gamma_2}, \hat{\mu_1}, \hat{\mu_2}) = (0.451, 0.549, 1.97, -1.04)$.

From Figure 2.3, we can see that it's more appropriate to use normal distributions with means 1.97 and -1.04 than standard normal distribution to describe the data. On the other hand, even if $\hat{\gamma_1}, \hat{\gamma_2}$ seem large, the means -1.04 and 1.97 are not particularly large. Thus, the fitted BCN model should not be taken to imply that there is meaningful differential expression on all genes of chromosome 21.

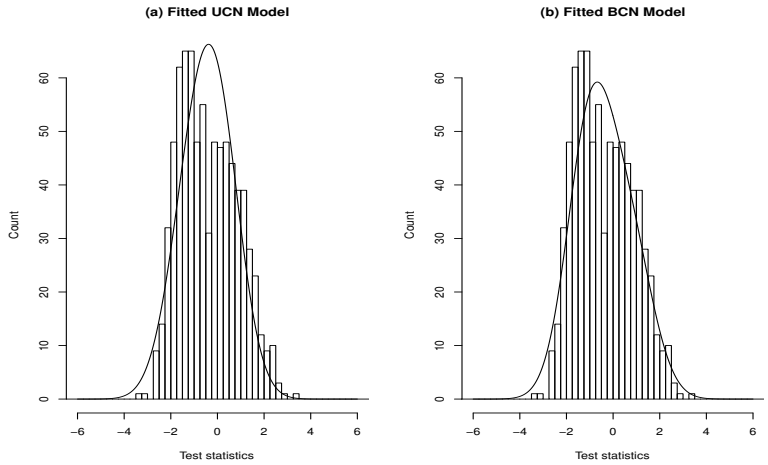

Figure 2.2: The fitted UCN and BCN models on chromosome 10.

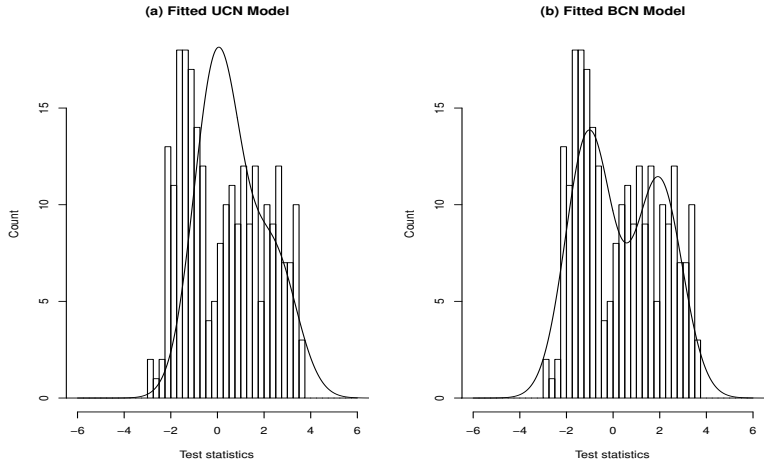

Figure 2.3: The fitted UCN and BCN models on chromosome 21.

Some of the test statistics are really large. Among 251 genes on chromosome 21, CXADR at chromosome location chr21q21.1, SOD1 at chr21q22.1/21q22.11, and PRMT2 at chr21q22.3 have test statistics exceeding 3.5. These three genes are identified as abnormal genes in patients who have Down's syndrome by Mao et al 2005, Wilcock 2012 and Blehaut et al 2010, respectively. We also note that while no gene on chromosome 21 had a test statistic below -3.5, gene CAPN5 on chromosome 11 did have a test statistic below -3.5; this was the only such gene.

Our methodology is different from Mao's in that Mao demonstrated dysregulation based on an evaluation of differential expression across many cell types (cerebrum, cerebellum, heart and astrocyte), while we focused on data obtained from cerebral tissue. Moreover, Mao et al didn't study the empirical distribution of test statistics, but they sought to identify genes that met a particular statistical criterion for dysregulation.

The goal of fitting the BCN model is to identify the presence of both over- and under-expression. For this purpose, BCN model recognized both types of differentially expressed genes in chromosome 21. And the fact that several genes have meaningful differential expression is confirmed by Figure 2.3.

Chapter 3 Bilaterally Contaminated Normal Model with Nuisance Parameter

3.1 Introduction

Since our previous work in Chapter 2 assumed a known within-component variance σ^2, in Chapter 3 we want to provide tests of contamination in the bilaterally contaminated normal model when the within-component variance σ^2 is unknown. The test for no contamination versus any contamination becomes

$$\begin{aligned} H_0 &: N(0,\sigma^2) \text{ versus} \\ H_a &: (1-\gamma_1-\gamma_2)N(0,\sigma^2)+\gamma_1 N(\mu_1,\sigma^2)+\gamma_2 N(\mu_2,\sigma^2) \neq N(0,\sigma^2) \end{aligned}$$

where $\gamma_1 \in [0, 1-\gamma_2]$, $\gamma_2 \in [0,1]$, $\mu_1 \geq 0$, $\mu_2 \leq 0$, and $\sigma^2 > 0$ are unknown but fixed parameters. A simple test based on moments is elusive. For example, consider defining a test statistic $T := \dfrac{3\hat{m}_2^2}{\hat{m}_4}$. This test statistic has a distribution that does not depend on σ^2 when H_0 is true. In fact, the asymptotic distribution when H_0 is true is given by

$$\sqrt{n}(T-1) \xrightarrow{L} N(0, \frac{8}{3}) \tag{3.1}$$

Proof. This is by the multivariate Central Limit Theorem (Ferguson, 1996, p.26) that under H_0, as $n \to \infty$,

$$\sqrt{n}\left[\begin{pmatrix} \hat{m}_2 \\ \hat{m}_4 \end{pmatrix} - \begin{pmatrix} \sigma^2 \\ 3\sigma^4 \end{pmatrix}\right] \xrightarrow{L} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2\sigma^4 & 12\sigma^6 \\ 12\sigma^6 & 96\sigma^8 \end{pmatrix}\right]. \tag{3.2}$$

Define $g(\boldsymbol{m}) = 3m_2^2/m_4$, then $g'(\boldsymbol{m}) = (6m_2/m_4, -3m_2^2/m_4^2) = (2/\sigma^2, -1/3\sigma^4)$.

By the multivariate Cramer Theorem (Ferguson, 1996, p.45),

$$\sqrt{n}(g(\hat{\boldsymbol{m}}) - g(\boldsymbol{m})) \xrightarrow{L} N(0, g'(\boldsymbol{m})\,\boldsymbol{\Sigma}\,g'(\boldsymbol{m})^T), \tag{3.3}$$

where $\mathbf{\Sigma}$ is the covariance matrix from (3.2). This gives us (3.1).

□

This suggests that we might reject H_0 if $|T-1|$ is too large, for instance greater than $Z_{1-\frac{\alpha}{2}}\sqrt{\frac{8}{3n}}$, where α is the desired significance level. Unfortunately, such a procedure is not consistent for all alternative hypotheses. Indeed, $\frac{3\hat{m}_2^2}{\hat{m}_4}$ may converge in probability to a number less than 1, to 1, or to a number greater than 1 under various alternative hypotheses. For example, when $\sigma^2 = 1, \mu_1 = 1, \mu_2 = -2, \gamma_1 = \gamma_2 = 0.05$, T converges in probability to 0.876 which is less than 1. If $\sigma^2 = 1, \mu_1 = 2, \mu_2 = -2, \gamma_1 = \gamma_2 = 1/6$, then T converges in probability to 1. But when $\sigma^2 = 1, \mu_1 = 1, \mu_2 = -1, \gamma_1 = \gamma_2 = 0.3$, T converges in probability to 1.067 which is larger than 1.

We therefore suggest to use a Union-Intersection test to solve the problem.

3.2 Omnibus Testing Procedure

The omnibus null hypothesis will be rejected when one of four conclusions is reached based on the data: (a) $m_1 \neq 0$, (b) $m_4/3m_2^2 \neq 1$, (c) $m_6/15m_2^3 \neq 1$, (d) $m_3 \neq 0$. The rationale for this is that at least one of these four conditions must hold when the omnibus null hypothesis is false (see proposition (3.2.5)). We will show how to achieve approximate significance level α for this *Union-Intersection* test, and we will prove that this *Union-Intersection* test is consistent against all alternatives.

Proposition 3.2.1. *If the omnibus null hypothesis is true, then* $T_n := \sqrt{n}\hat{m}_1$ *converges in law to* $N(0, \sigma^2)$. *Thus* $\sqrt{\frac{n}{\hat{\sigma}^2}}\hat{m}_1$ *converges in law to* $N(0,1)$.

Proof. By CLT, $\sqrt{n}(\hat{m}_1 - 0) \xrightarrow{L} N(0, Var(X_1))$ under the omnibus null, where $Var(X_1) = \sigma^2$. Then apply Slutsky's Theorem. □

Proposition 3.2.2. *If the omnibus null hypothesis is true, then* $U_n := \sqrt{n}(\hat{m}_4/3\hat{m}_2^2 - 1)$ *converges in law to* $N(0, 8/3)$. *Thus* $\sqrt{\frac{3n}{8}}(\hat{m}_4/3\hat{m}_2^2 - 1)$ *converges in law to* $N(0, 1)$.

Proof. By (3.1). □

Proposition 3.2.3. *If the omnibus null hypothesis is true, then* $V_n := \sqrt{n}(\hat{m}_6/15\hat{m}_2^3 - 1)$ *converges in law to* $N(0, 136/5)$. *Thus* $\sqrt{\frac{5n}{136}}(\hat{m}_6/15\hat{m}_2^3 - 1)$ *converges in law to* $N(0, 1)$.

Proof. Similar to the proof of previous proposition. □

Proposition 3.2.4. *If the omnibus null hypothesis is true, then* $W_n := \sqrt{n}\hat{m}_3$ *converges in law to* $N(0, 15\sigma^6)$. *Thus* $\sqrt{\frac{n}{15\hat{\sigma}^6}}\hat{m}_1$ *converges in law to* $N(0, 1)$.

Proof. By CLT, $\sqrt{n}(\hat{m}_3 - 0) \xrightarrow{L} N(0, Var(X_1^3))$ under the omnibus null, where $Var(X_1^3) = 15\sigma^6$. Then apply Slutsky's Theorem. □

Proposition 3.2.5. *If the omnibus null hypothesis is false, then at least one of the following conditions must hold: (a)* $m_1 \neq 0$, *(b)* $m_4/3m_2^2 \neq 1$, *(c)* $m_6/15m_2^3 \neq 1$, *(d)* $m_3 \neq 0$.

Proof. Now assume the omnibus null is false. First consider symmetric case, assume $\mu_1 = -\mu_2 = \mu \geq 0$, $\gamma_1 = \gamma_2 = \gamma$, then the BCN+NP model can be written as

$$(1 - 2\gamma)N(0, \sigma^2) + \gamma N(\mu, \sigma^2) + \gamma N(-\mu, \sigma^2).$$

The second, fourth and sixth moments become

$m_2 = \sigma^2 + 2\gamma\mu^2$

$m_4 = 3\sigma^4 + 2\gamma(6\sigma^2\mu^2 + \mu^4)$

$m_6 = 15\sigma^6 + 2\gamma(45\sigma^4\mu^2 + 15\sigma^2\mu^4 + \mu^6)$

Under this case, either $m_4/3m_2^2 = 1$ or $m_4/3m_2^2 \neq 1$. If $m_4/3m_2^2 = 1$, an easy calculation can show that $\gamma = 1/6$. Then if $\mu > 0$, $m_6/15m_2^3 \neq 1$. And if $\mu = 0$, we

contradict the assumption that H_a holds.

Next consider the case of asymmetry. In this case, we can consider $m_1 = 0$ and $m_1 \neq 0$. If $m_1 = 0$, this means $\mu_1\gamma_1 + \mu_2\gamma_2 = 0$. Then $m_3 = \gamma_2\mu_2(\mu_2^2 - \mu_1^2)$. If $\mu_1 \neq \mu_2$, $m_3 \neq 0$. If $\mu_1 = \mu_2$, we contradict the assumption that H_a holds.

□

The above propositions yield the following theorem,

Theorem 3.2.6. *Let α_1, α_2, α_3, $\alpha_4 \in (0,1)$, and $\sum_{i=1}^{4} \alpha_i = \alpha \in (0,1)$. Consider a test that rejects the omnibus null hypothesis if at least one of the following four conditions holds: (a) $|T_n|/\hat{\sigma} > Z_{1-\alpha_1/2}$, (b) $\sqrt{3/8}|U_n| > Z_{1-\alpha_2/2}$, (c) $\sqrt{5/136}|V_n| > Z_{1-\alpha_3/2}$, (d) $|W_n|/\sqrt{15}\hat{\sigma}^3 > Z_{1-\alpha_4/2}$. This test rejects a true omnibus null hypothesis with probability less than or equal to α, as $n \to \infty$, and rejects a false omnibus null hypothesis with (approximate) probability greater than or equal to*

$$\Phi\left[\max\left(-Z_{1-\frac{\alpha_1}{2}} + \sqrt{\frac{n}{\sigma^2}}|m_1|,\ -Z_{1-\frac{\alpha_2}{2}} + \sqrt{\frac{3n}{8}}\left|\frac{m_4}{3m_2^2} - 1\right|,\right.\right.$$

$$\left.\left.-Z_{1-\frac{\alpha_3}{2}} + \sqrt{\frac{5n}{136}}\left|\frac{m_6}{15m_2^3} - 1\right|,\ -Z_{1-\frac{\alpha_4}{2}} + \sqrt{\frac{n}{15\sigma^6}}|m_3|\right)\right],$$

where Φ denotes the standard normal cumulative distribution function. Hence, this test procedure is consistent.

Proof. $P((a)\bigcup(b)\bigcup(c)\bigcup(d)) \leq P((a)) + P((b)) + P((c)) + P((d)) \to \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = \alpha$ when the omnibus null hypothesis is true by *Proposition* (3.2.1) through

(3.2.4) respectively. Now suppose the omnibus null hypothesis is false. We have

$$
\begin{aligned}
P\left(\sqrt{\frac{5}{136}}\,|V_n| > Z_{1-\frac{\alpha_3}{2}}\right) &= P\left(\sqrt{\frac{5}{136}}V_n > Z_{1-\frac{\alpha_3}{2}}\right) + P\left(\sqrt{\frac{5}{136}}V_n < -Z_{1-\frac{\alpha_3}{2}}\right) \\
&= P\left(\sqrt{\frac{5n}{136}}\left(\frac{\hat{m}_6}{15\hat{m}_2^3} - 1\right) > Z_{1-\frac{\alpha_3}{2}}\right) + P\left(\sqrt{\frac{5n}{136}}\left(\frac{\hat{m}_6}{15\hat{m}_2^3} - 1\right) < -Z_{1-\frac{\alpha_3}{2}}\right) \\
&= P\left(\sqrt{\frac{5n}{136}}\left(\frac{\hat{m}_6}{15\hat{m}_2^3} - \frac{m_6}{15m_2^3}\right) > Z_{1-\frac{\alpha_3}{2}} + \sqrt{\frac{5n}{136}}\left(1 - \frac{m_6}{15m_2^3}\right)\right) + \\
&\quad P\left(\sqrt{\frac{5n}{136}}\left(\frac{\hat{m}_6}{15\hat{m}_2^3} - \frac{m_6}{15m_2^3}\right) < -Z_{1-\frac{\alpha_3}{2}} + \sqrt{\frac{5n}{136}}\left(1 - \frac{m_6}{15m_2^3}\right)\right) \\
&\approx 1 - \Phi\left(Z_{1-\frac{\alpha_3}{2}} + \sqrt{\frac{5n}{136}}\left(1 - \frac{m_6}{15m_2^3}\right)\right) + \Phi\left(-Z_{1-\frac{\alpha_3}{2}} + \sqrt{\frac{5n}{136}}\left(1 - \frac{m_6}{15m_2^3}\right)\right) \\
&= \Phi\left(-Z_{1-\frac{\alpha_3}{2}} - \sqrt{\frac{5n}{136}}\left(1 - \frac{m_6}{15m_2^3}\right)\right) + \Phi\left(-Z_{1-\frac{\alpha_3}{2}} + \sqrt{\frac{5n}{136}}\left(1 - \frac{m_6}{15m_2^3}\right)\right)
\end{aligned}
$$

This is approximately equal to $\Phi\left(-Z_{1-\frac{\alpha_3}{2}} - \sqrt{\frac{5n}{136}}\left(1 - \frac{m_6}{15m_2^3}\right)\right)$ if $\frac{m_6}{15m_2^3} > 1$ and to $\Phi\left(-Z_{1-\frac{\alpha_3}{2}} + \sqrt{\frac{5n}{136}}\left(1 - \frac{m_6}{15m_2^3}\right)\right)$ if $\frac{m_6}{15m_2^3} < 1$. So $P\left(\sqrt{\frac{5}{136}}|V_n| > Z_{1-\frac{\alpha_3}{2}}\right) \approx \Phi\left(-Z_{1-\frac{\alpha_3}{2}} + \sqrt{\frac{5n}{136}}\left|\frac{m_6}{15m_2^3} - 1\right|\right)$. Similarly, $P\left(\frac{3}{8}|U_n| > Z_{1-\frac{\alpha_2}{2}}\right) \approx \Phi\left(-Z_{1-\frac{\alpha_2}{2}} + \sqrt{\frac{3n}{8}}\left|\frac{m_4}{15m_2^2} - 1\right|\right)$.
Also,

$$
\begin{aligned}
P\left(\frac{|W_n|}{\sqrt{15\hat{\sigma}^6}} > Z_{1-\frac{\alpha_4}{2}}\right) &= P\left(\frac{W_n}{\sqrt{15\hat{\sigma}^6}} > Z_{1-\frac{\alpha_4}{2}}\right) + P\left(\frac{W_n}{\sqrt{15\hat{\sigma}^6}} < -Z_{1-\frac{\alpha_4}{2}}\right) \\
&= P\left(\sqrt{\frac{n}{15\hat{\sigma}^6}}\hat{m}_3 > Z_{1-\frac{\alpha_4}{2}}\right) + P\left(\sqrt{\frac{n}{15\hat{\sigma}^6}}\hat{m}_3 < -Z_{1-\frac{\alpha_4}{2}}\right) \\
&= P\left(\sqrt{\frac{n}{15\hat{\sigma}^6}}\left(\hat{m}_3 - m_3\right) > Z_{1-\frac{\alpha_4}{2}} - \sqrt{\frac{n}{15\hat{\sigma}^6}}m_3\right) + \\
&\quad P\left(\sqrt{\frac{n}{15\hat{\sigma}^6}}\left(\hat{m}_3 - m_3\right) < -Z_{1-\frac{\alpha_4}{2}} - \sqrt{\frac{n}{15\hat{\sigma}^6}}m_3\right) \\
&\approx 1 - \Phi\left(Z_{1-\frac{\alpha_4}{2}} - \sqrt{\frac{n}{15\sigma^6}}m_3\right) + \Phi\left(-Z_{1-\frac{\alpha_4}{2}} - \sqrt{\frac{n}{15\sigma^6}}m_3\right) \\
&= \Phi\left(-Z_{1-\frac{\alpha_4}{2}} + \sqrt{\frac{n}{15\sigma^6}}m_3\right) + \Phi\left(-Z_{1-\frac{\alpha_4}{2}} - \sqrt{\frac{n}{15\sigma^6}}m_3\right)
\end{aligned}
$$

This is approximately equal to $\Phi\left(-Z_{1-\frac{\alpha_4}{2}} + \sqrt{\frac{n}{15\sigma^6}}m_3\right)$ if $m_3 > 0$ and $\Phi\left(-Z_{1-\frac{\alpha_4}{2}} - \sqrt{\frac{n}{15\sigma^6}}m_3\right)$ if $m_3 < 0$. So $P\left(\frac{|W_n|}{\sqrt{15\hat{\sigma}^6}} > Z_{1-\frac{\alpha_4}{2}}\right) \approx \Phi\left(-Z_{1-\frac{\alpha_4}{2}} + \sqrt{\frac{n}{15\sigma^6}}|m_3|\right)$. Similarly, one can

show that $P\left(\frac{|T_n|}{\sqrt{\hat{\sigma}^2}} > Z_{1-\frac{\alpha_1}{2}}\right) \approx \Phi\left(-Z_{1-\frac{\alpha_1}{2}} + \sqrt{\frac{n}{\sigma^2}}|m_1|\right)$.

Now since $P((a) \bigcup (b) \bigcup (c) \bigcup (d)) \geq P((a)), P((b)), P((c))$ and $P((d))$ respectively, the probability of the union is greater than or equal to $\max P((a)), P((b)), P((c)), P((d))$.

$$\lim_{n\to\infty} P((a) \bigcup (b) \bigcup (c) \bigcup (d)) \geq \lim_{n\to\infty} \Phi\left[\max\left(-Z_{1-\frac{\alpha_1}{2}} + \sqrt{\frac{n}{\sigma^2}}|m_1|,\right.\right.$$
$$- Z_{1-\frac{\alpha_2}{2}} + \sqrt{\frac{3n}{8}}\left|\frac{m_4}{3m_2^2} - 1\right|,\ -Z_{1-\frac{\alpha_3}{2}} + \sqrt{\frac{5n}{136}}\left|\frac{m_6}{15m_2^3} - 1\right|,$$
$$\left.\left.-Z_{1-\frac{\alpha_4}{2}} + \sqrt{\frac{n}{15\sigma^6}}|m_3|\right)\right] = 1.$$

Hence the test is consistent.

□

3.3 Unilateral Testing Procedure

The test for unilateral contamination versus bilateral contamination with σ^2 unknown becomes

$$\begin{aligned} H_0 &: (1-\gamma)N(0,\sigma^2) + \gamma N(\mu,\sigma^2) \neq N(0,\sigma^2) \text{ versus} \\ H_a &: (1-\gamma_1-\gamma_2)N(0,\sigma^2) + \gamma_1 N(\mu_1,\sigma^2) + \gamma_2 N(\mu_2,\sigma^2) \neq N(0,\sigma^2) \\ &\quad \text{and } \neq (1-\gamma)N(0,\sigma^2) + \gamma N(\mu,\sigma^2) \end{aligned}$$

where $\gamma_1 \in [0, 1-\gamma_2]$, γ, $\gamma_2 \in [0,1]$, μ, $\mu_1 \geq 0$, $\mu_2 \leq 0$, and $\sigma^2 > 0$ are unknown but fixed parameters. An easy test for this procedure is not available due to the complexity of the unilateral null hypothesis. However, a moment based test, though complex, is available. We will show this test rejects the unilateral null with probability asymptotically bounded above by α under H_0 and with probability approaching 1 under H_a.

Lemma 3.3.1. *Under the unilateral null and assuming $\gamma < 2/3$, we have*

$$\sigma^2 = m_2 - \frac{1}{2}\left(3m_1^2 + \sqrt{9m_1^4 - 12m_1^2m_2 + 4m_1m_3}\right).$$

Proof. Assume the unilateral null is true, that is

$$(1-\gamma)N(0,\sigma^2) + \gamma N(\mu,\sigma^2).$$

The first three moments become

$$m_1 = \mu\gamma,$$

$$m_2 = \sigma^2 + \gamma\mu^2,$$

$$m_3 = \gamma\mu^3 + 3\gamma\sigma^2\mu.$$

Solving the equations for μ, σ^2 and γ yields

$$\mu = \frac{1}{2m_1}\left(3m_1^2 + \sqrt{9m_1^4 - 12m_1^2m_2 + 4m_1m_3}\right),$$

$$\sigma^2 = m_2 - 0.5\left(3m_1^2 + \sqrt{9m_1^4 - 12m_1^2m_2 + 4m_1m_3}\right),$$

$$\gamma = 2m_1^2 / \left(3m_1^2 + \sqrt{9m_1^4 - 12m_1^2m_2 + 4m_1m_3}\right).$$

The assumption that $\gamma < 2/3$ leads to selection of the positive square root when applying the quadratic formula. □

Lemma 3.3.2. *Under the unilateral null, assume $\gamma < 2/3$ and that the omnibus null is false. Then*

$$\sqrt{n}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{L} N(0, \boldsymbol{D\Sigma D^T}),$$

where $\boldsymbol{D}$ is the vector of partial derivatives of $\sigma^2 = m_2 - 0.5\left(3m_1^2 + \sqrt{9m_1^4 - 12m_1^2m_2 + 4m_1m_3}\right)$ with respect to m_1, m_2, m_3.

Proof. The third component of $\boldsymbol{D} = \dfrac{-m_1}{\sqrt{9m_1^4 - 12m_1^2m_2 + 4m_1m_3}} \neq 0$, since $m_1 = \mu\gamma \neq 0$ because the omnibus null is false. So $\boldsymbol{\Sigma}$ is positive definite implies $\boldsymbol{D\Sigma D^T}$ is

positive. By the multivariate Central Limit Theorem, we have

$$\sqrt{n}\left[\begin{pmatrix}\hat{m}_1\\ \hat{m}_2\\ \hat{m}_3\end{pmatrix}-\begin{pmatrix}m_1\\ m_2\\ m_3\end{pmatrix}\right]\xrightarrow{L} N\left[\begin{pmatrix}0\\ 0\\ 0\end{pmatrix},\ \mathbf{\Sigma}\right], \tag{3.4}$$

where each component of matrix $\mathbf{\Sigma}$ is $\Sigma_{ij} = m_{i+j} - m_i m_j$, $i = 1, 2, 3$, $j = 1, 2, 3$. By multivariate Delta Method, the above result holds. □

Proposition 3.3.3. *Suppose unilateral null is true and omnibus null is false. If* $|\hat{\sigma}^2 - \sigma^2| \leq \delta_n$ *for some positive* δ_n*, then with probability approaching 1 we have*

$$\begin{aligned}&(\hat{m}_2 - \sigma^2)^2 - 2\delta_n\hat{m}_2 + 3\hat{m}_1^2\sigma^2 - 3\hat{m}_1^2\delta_n - \hat{m}_1\hat{m}_3\\ &\qquad\leq (\hat{m}_2 - \hat{\sigma}^2)^2 + 3\hat{m}_1^2\hat{\sigma}^2 - \hat{m}_1\hat{m}_3\\ &\qquad\leq (\hat{m}_2 - \sigma^2)^2 + \delta_n^2 + 2\delta_n\hat{m}_2 + 3\hat{m}_1^2\sigma^2 + 3\hat{m}_1^2\delta_n - \hat{m}_1\hat{m}_3.\end{aligned}$$

Proof. We have $(\hat{m}_2 - \hat{\sigma}^2)^2 = (\hat{m}_2 - \sigma^2)^2 + (\sigma^2 - \hat{\sigma}^2)^2 + 2(\hat{m}_2 - \sigma^2)(\sigma^2 - \hat{\sigma}^2)$
Under unilateral null with false omnibus null, $m_2 > \sigma^2 + m_1^2 > \sigma^2$, so by Weak Law of Large Numbers $\hat{m}_2 \xrightarrow{P} m_2 > \sigma^2$ and $P(\hat{m}_2 - \sigma^2 > 0) \to 1$. So following statements hold with probability approaching 1:

$$(\hat{m}_2 - \sigma^2)^2 - 2\delta_n|\hat{m}_2 - \sigma^2| \leq (\hat{m}_2 - \hat{\sigma}^2)^2 \leq (\hat{m}_2 - \sigma^2)^2 + \delta_n^2 + 2\delta_n|\hat{m}_2 - \sigma^2|$$

$$(\hat{m}_2 - \sigma^2)^2 - 2\delta_n\hat{m}_2 + 2\delta_n\sigma^2 \leq (\hat{m}_2 - \hat{\sigma}^2)^2 \leq (\hat{m}_2 - \sigma^2)^2 + \delta_n^2 + 2\delta_n\hat{m}_2 - 2\delta_n\sigma^2$$

$$(\hat{m}_2 - \sigma^2)^2 - 2\delta_n\hat{m}_2 \leq (\hat{m}_2 - \hat{\sigma}^2)^2 \leq (\hat{m}_2 - \sigma^2)^2 + \delta_n^2 + 2\delta_n\hat{m}_2 \tag{3.5}$$

And

$$3\hat{m}_1^2(\sigma^2 - \delta_n) - \hat{m}_1\hat{m}_3 \leq 3\hat{m}_1^2\hat{\sigma}^2 - \hat{m}_1\hat{m}_3 \leq 3\hat{m}_1^2(\sigma^2 + \delta_n) - \hat{m}_1\hat{m}_3 \tag{3.6}$$

Thus the desired result holds by addition of (3.5) and (3.6). □

Now define $\boldsymbol{m} = (m_1, m_2, m_3)$, $\hat{\boldsymbol{m}} = (\hat{m}_1, \hat{m}_2, \hat{m}_3)$, $h(\boldsymbol{m}, \sigma^2) := (m_2 - \sigma^2)^2 + 3m_1^2\sigma^2 - m_1 m_3$.

Proposition 3.3.4. *Suppose unilateral null is true and omnibus null is false. Then* $h(\boldsymbol{m}, \sigma^2) = 0$ *and thus* $P\left(h(\hat{\boldsymbol{m}}, \sigma^2) > Z_{1-\alpha}\sqrt{\frac{\hat{\boldsymbol{D}}_h \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{D}}_h^T}{n}}\right) \to 1 - \Phi(Z_{1-\alpha}) = \alpha$, *where* $\hat{\boldsymbol{D}}_{\boldsymbol{h}} = \frac{\partial}{\partial \hat{\boldsymbol{m}}} h(\hat{\boldsymbol{m}}, \sigma^2) = (6\hat{m}_1\sigma^2, 2\hat{m}_2 - 2\sigma^2, -\hat{m}_1)$.

Proof. Note that $\boldsymbol{D}_{\boldsymbol{h}} = (6m_1\sigma^2, 2m_2 - 2\sigma^2, -m_1) \neq \boldsymbol{0}$ since $m_1 \neq 0$ when unilateral null is true and omnibus null is false.

By multivariate Delta Method,

$$\sqrt{n}(h(\hat{\boldsymbol{m}}, \sigma^2) - h(\boldsymbol{m}, \sigma^2)) \xrightarrow{d} N(0, \boldsymbol{D}_{\boldsymbol{h}} \boldsymbol{\Sigma} \boldsymbol{D}_{\boldsymbol{h}}^{\boldsymbol{T}}).$$

Thus by definition of convergence in distribution,

$$P\left(\sqrt{\frac{n}{\boldsymbol{D}_{\boldsymbol{h}} \boldsymbol{\Sigma} \boldsymbol{D}_{\boldsymbol{h}}^{\boldsymbol{T}}}}(h(\hat{\boldsymbol{m}}, \sigma^2) - h(\boldsymbol{m}, \sigma^2)) > Z_{1-\alpha}\right) \to 1 - \Phi(Z_{1-\alpha})$$

.

We have $(\hat{m}_1, \hat{m}_2, \hat{m}_3)^T \xrightarrow{P} (m_1, m_2, m_3)^T$ by the Weak Law of Large Numbers. Then by Slutsky's Theorem and Continuous Mapping Theorem, $\hat{\boldsymbol{D}}_{\boldsymbol{h}} \xrightarrow{P} \boldsymbol{D}_{\boldsymbol{h}}$ and

$$P\left(\sqrt{\frac{n}{\hat{\boldsymbol{D}}_{\boldsymbol{h}} \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{D}}_{\boldsymbol{h}}^{\boldsymbol{T}}}}(h(\hat{\boldsymbol{m}}, \sigma^2) - h(\boldsymbol{m}, \sigma^2)) > Z_{1-\alpha}\right) \to 1 - \Phi(Z_{1-\alpha}).$$

Thus,

$$P\left(h(\hat{\boldsymbol{m}}, \sigma^2) > h(\boldsymbol{m}, \sigma^2) + Z_{1-\alpha}\sqrt{\frac{\hat{\boldsymbol{D}}_{\boldsymbol{h}} \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{D}}_{\boldsymbol{h}}^{\boldsymbol{T}}}{n}}\right) \to 1 - \Phi(Z_{1-\alpha}). \tag{3.7}$$

Under unilateral null, $h(\boldsymbol{m}, \sigma^2) = \mu_1|\mu_2|\gamma_1\gamma_2(\mu_1 - \mu_2)^2 = 0$, then the desired result holds. □

Theorem 3.3.5. *Let* δ_n *be chosen so that* $P(|\hat{\sigma}^2 - \sigma^2| \leq \delta_n) \to 1$. *Consider the testing procedure defined by rejecting the null hypothesis when* $h(\hat{\boldsymbol{m}}, \hat{\sigma}^2) > Z_{1-\alpha}\sqrt{\frac{\hat{\boldsymbol{D}}_h \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{D}}_h^T}{n}} + 2\hat{m}_2\delta_n + 3\hat{m}_1\delta_n + \delta_n^2$. *Type I error probability asymptotically bounded above by* α *for any fixed* $\mu_1, \mu_2, \gamma_1, \gamma_2$, *such that* $\mu_1\mu_2\gamma_1\gamma_2 = 0$ *with* $\mu_1\gamma_1 \neq 0$ *or* $\mu_2\gamma_2 \neq 0$.

Proof. By (3.3.3), $h(\hat{\boldsymbol{m}}, \hat{\sigma}^2) \leq h(\hat{\boldsymbol{m}}, \sigma^2) + 2\hat{m}_2\delta_n + 3\hat{m}_1\delta_n + \delta_n^2$ given $P(|\hat{\sigma}^2 - \sigma^2| \leq \delta_n) \to 1$. If in addition assume $h(\hat{\boldsymbol{m}}, \sigma^2) \leq Z_{1-\alpha}\sqrt{\frac{\hat{\boldsymbol{D}}_{\boldsymbol{h}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{D}}_{\boldsymbol{h}}^{\boldsymbol{T}}}{n}}$, we have $h(\hat{\boldsymbol{m}}, \hat{\sigma}^2) \leq Z_{1-\alpha}\sqrt{\frac{\hat{\boldsymbol{D}}_{\boldsymbol{h}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{D}}_{\boldsymbol{h}}^{\boldsymbol{T}}}{n}} + 2\hat{m}_2\delta_n + 3\hat{m}_1\delta_n + \delta_n^2$. Thus

$$
\begin{aligned}
& P\left(h(\hat{\boldsymbol{m}}, \hat{\sigma}^2) > Z_{1-\alpha}\sqrt{\frac{\hat{\boldsymbol{D}}_{\boldsymbol{h}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{D}}_{\boldsymbol{h}}^{\boldsymbol{T}}}{n}} + 2\hat{m}_2\delta_n + 3\hat{m}_1\delta_n + \delta_n^2\right) \\
& \leq P\left(|\hat{\sigma}^2 - \sigma^2| > \delta_n \text{ or } h(\hat{\boldsymbol{m}}, \sigma^2) > Z_{1-\alpha}\sqrt{\frac{\hat{\boldsymbol{D}}_{\boldsymbol{h}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{D}}_{\boldsymbol{h}}^{\boldsymbol{T}}}{n}}\right) \\
& \leq P\left(|\hat{\sigma}^2 - \sigma^2| > \delta_n\right) + P\left(h(\hat{\boldsymbol{m}}, \sigma^2) > Z_{1-\alpha}\sqrt{\frac{\hat{\boldsymbol{D}}_{\boldsymbol{h}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{D}}_{\boldsymbol{h}}^{\boldsymbol{T}}}{n}}\right) \\
& \to \alpha
\end{aligned}
$$

by assumption and *Proposition* (3.3.4). □

Corollary 3.3.6. *Let $\delta_n \propto n^{-1/2+\epsilon}$ for arbitrary $\epsilon \in (0, 1/2)$. Then for fixed μ_1, μ_2, γ_1, γ_2, σ^2, Type I error probability of testing procedure converges to α.*

Proof. If $\delta_n \propto n^{-1/2+\epsilon}$, then by Lemma (3.3.2). $P(|\hat{\sigma}^2 - \sigma^2| \leq \delta_n) \to 1$. Apply Theorem (3.3.5) □

Proposition 3.3.7. *Suppose the unilateral null hypothesis is false, then*

$$\sqrt{n}(h(\hat{\boldsymbol{m}}, \sigma^2) - h(\boldsymbol{m}, \sigma^2)) \xrightarrow{L} N(0, \boldsymbol{D}_{\boldsymbol{h}}\boldsymbol{\Sigma}\boldsymbol{D}_{\boldsymbol{h}}^{\boldsymbol{T}}),$$

where $\boldsymbol{\Sigma}$ is defined in (3.4)*, $\boldsymbol{D}_{\boldsymbol{h}}$ is the vector of partial derivatives of $h(\boldsymbol{m}, \sigma^2) = (m_2 - \sigma^2)^2 + 3m_1^2\sigma^2 - m_1m_3 = \mu_1|\mu_2|\gamma_1\gamma_2(\mu_1 - \mu_2)^2$ with respect to $\boldsymbol{m}$.*

Proof. Under bilateral alternative hypothesis, $h(\boldsymbol{m}, \sigma^2) = \mu_1|\mu_2|\gamma_1\gamma_2(\mu_1 - \mu_2)^2 > 0$. And since the second component of $\boldsymbol{D}_{\boldsymbol{h}}$ is $2(m_2 - \sigma^2) = 2(\mu_1^2\gamma_1 - \mu_2^2\gamma_2) > 0$, $\boldsymbol{D}_{\boldsymbol{h}} \neq \boldsymbol{0}$. On the other hand, $h(\hat{\boldsymbol{m}}, \sigma^2) = (\hat{m}_2 - \sigma^2)^2 + 3\hat{m}_1^2\sigma^2 - \hat{m}_1\hat{m}_3$, then by multivariate Delta method,

$$\sqrt{n}(h(\hat{\boldsymbol{m}}, \sigma^2) - h(\boldsymbol{m}, \sigma^2)) \xrightarrow{L} N(0, \boldsymbol{D}_{\boldsymbol{h}}\boldsymbol{\Sigma}\boldsymbol{D}_{\boldsymbol{h}}^{\boldsymbol{T}}).$$

□

Proposition 3.3.8. *Suppose the unilateral null hypothesis is false, then*

$$P\left(h(\hat{\boldsymbol{m}},\sigma^2) > h(\boldsymbol{m},\sigma^2) + Z_\alpha\sqrt{\frac{\hat{\boldsymbol{D}}_{\boldsymbol{h}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{D}}_{\boldsymbol{h}}^{\boldsymbol{T}}}{n}}\right) \to 1-\Phi(Z_{1-\alpha}) = \alpha,$$

where $\hat{\boldsymbol{D}}_{\boldsymbol{h}} = \frac{\partial}{\partial\hat{\boldsymbol{m}}}h(\hat{\boldsymbol{m}},\sigma^2) = (6\hat{m}_1\sigma^2, 2\hat{m}_2 - 2\sigma^2, -\hat{m}_1)$.

Proof. Note that $\boldsymbol{D}_{\boldsymbol{h}} = (6m_1\sigma^2, 2m_2 - 2\sigma^2, -m_1) \neq \boldsymbol{0}$, since $m_2 \neq \sigma^2$ under the bilateral alternative. Then the result follows from essentially the same argument used to establish (3.7). □

Theorem 3.3.9. *Let fixed* $\delta_n > 0$ *be chosen so that* $P(|\hat{\sigma}^2 - \sigma^2| \leq \delta_n) \to 1$, $P(|\hat{m}_1^2 - m_1^2| \leq \delta_n) \to 1$ *and* $P(|\hat{m}_2 - m_2| \leq \delta_n) \to 1$ *as* $n \to \infty$. *Consider the testing procedure defined by rejecting the unilateral null hypothesis when* $h(\hat{\boldsymbol{m}},\hat{\sigma}^2) > 2\hat{m}_2\delta_n + 3\hat{m}_1^2\delta_n + \delta_n^2 + Z_{1-\alpha}\sqrt{\frac{\hat{\boldsymbol{D}}_{\boldsymbol{h}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{D}}_{\boldsymbol{h}}^{\boldsymbol{T}}}{n}}$. *The power of this test has an approximate lower bound of* $1-\Phi\left(Z_{1-\alpha} + \sqrt{\frac{n}{\boldsymbol{D}_{\boldsymbol{h}}\boldsymbol{\Sigma}\boldsymbol{D}_{\boldsymbol{h}}^{\boldsymbol{T}}}}\left(4m_2\delta_n + 6m_1^2\delta_n + 11\delta_n^2 - \mu_1|\mu_2|\gamma_1\gamma_2(\mu_1-\mu_2)^2\right)\right)$. *Moreover, this test is consistent whenever* $4m_2\delta_n + 6m_1^2\delta_n + 11\delta_n^2 < h(\boldsymbol{m},\sigma^2)$.

Proof. By *Proposition* (3.3.3),

$$\begin{aligned}
&P\left(h(\hat{\boldsymbol{m}},\hat{\sigma}^2) > 2\hat{m}_2\delta_n + 3\hat{m}_1^2\delta_n + \delta_n^2 + Z_{1-\alpha}\sqrt{\frac{\hat{\boldsymbol{D}}_{\boldsymbol{h}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{D}}_{\boldsymbol{h}}^{\boldsymbol{T}}}{n}}\right) \\
\leq\ &P\left(h(\hat{\boldsymbol{m}},\sigma^2) - 2\hat{m}_2\delta_n - 3\hat{m}_1^2\delta_n > 2\hat{m}_2\delta_n + 3\hat{m}_1^2\delta_n + \delta_n^2 + Z_{1-\alpha}\sqrt{\frac{\hat{\boldsymbol{D}}_{\boldsymbol{h}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{D}}_{\boldsymbol{h}}^{\boldsymbol{T}}}{n}}\right) - \\
&\quad P(|\hat{\sigma}^2 - \sigma^2| > \delta_n)
\end{aligned}$$

$$\leq P\left(h(\hat{\boldsymbol{m}},\sigma^2) > 4m_2\delta_n + 6m_1^2\delta_n + 11\delta_n^2 + Z_{1-\alpha}\sqrt{\frac{\hat{\boldsymbol{D}}_{\boldsymbol{h}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{D}}_{\boldsymbol{h}}^{\boldsymbol{T}}}{n}}\right) -$$

$$P(|\hat{\sigma}^2 - \sigma^2| > \delta_n) - P(|\hat{m}_1^2 - m_1^2| > \delta_n) - P(|\hat{m}_2 - m_2| > \delta_n)$$

$$= P\left(\sqrt{n}\left(h\left(\hat{\boldsymbol{m}},\sigma^2\right) - h\left(\boldsymbol{m},\sigma^2\right)\right) > \sqrt{n}\left(4m_2\delta_n + 6m_1^2\delta_n + 11\delta_n^2 + Z_{1-\alpha}\sqrt{\frac{\hat{\boldsymbol{D}}_{\boldsymbol{h}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{D}}_{\boldsymbol{h}}^{\boldsymbol{T}}}{n}} - \right.\right.$$

$$\left.\left. h\left(\boldsymbol{m},\sigma^2\right)\right)\right) - P(|\hat{\sigma}^2 - \sigma^2| > \delta_n) - P(|\hat{m}_1^2 - m_1^2| > \delta_n) - P(|\hat{m}_2 - m_2| > \delta_n)$$

$$\approx 1 - \Phi\left(Z_{1-\alpha} + \sqrt{\frac{n}{\boldsymbol{D}_{\boldsymbol{h}}\boldsymbol{\Sigma}\boldsymbol{D}_{\boldsymbol{h}}^{\boldsymbol{T}}}}\left(4m_2\delta_n + 6m_1^2\delta_n + 11\delta_n^2 - \mu_1|\mu_2|\gamma_1\gamma_2(\mu_1 - \mu_2)^2\right)\right)$$

The above term converges to 1 if $4m_2\delta_n + 6m_1^2\delta_n + 11\delta_n^2 < \mu_1|\mu_2|\gamma_1\gamma_2(\mu_1 - \mu_2)^2$.

□

Corollary 3.3.10. *Suppose* $\delta_n \to 0$ *with* $P(|\hat{\sigma}^2 - \sigma^2| \leq \delta_n) \to 1$, $P(|\hat{m}_1^2 - m_1^2| \leq \delta_n) \to 1$ *and* $P(|\hat{m}_2 - m_2| \leq \delta_n) \to 1$, *as* $n \to \infty$, *then the test in Theorem* (3.3.9) *is consistent against any fixed alternative.*

Proof. For any fixed alternative, we may find fixed $\delta^* > 0$ such that $4m_2\delta^* + 6m_1^2\delta^* + 11\delta^{*2} < h(\boldsymbol{m},\sigma^2)$. For large enough n, $\delta_n < \delta^*$ since $\delta_n \to 0$. Thus $P(h(\hat{\boldsymbol{m}},\sigma^2) > 4m_2\delta_n + 6m_1^2\delta_n + 11\delta_n^2 + Z_{1-\alpha}\sqrt{\frac{\hat{\boldsymbol{D}}_{\boldsymbol{h}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{D}}_{\boldsymbol{h}}^{\boldsymbol{T}}}{n}}) \geq P(h(\hat{\boldsymbol{m}},\sigma^2) > 4m_2\delta^* + 6m_1^2\delta^* + 11\delta^{*2} + Z_{1-\alpha}\sqrt{\frac{\hat{\boldsymbol{D}}_{\boldsymbol{h}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{D}}_{\boldsymbol{h}}^{\boldsymbol{T}}}{n}}) \to 1$ by Theorem (3.3.9). Hence the test is consistent against the fixed alternative. □

As proved in the previous theorems, the testing procedure for omnibus null hypothesis against bilateral alternative hypothesis and the testing procedure for unilateral null hypothesis against bilateral alternative hypothesis are both consistent. The bilaterally contaminated normal model with nuisance parameter is more flexible than the bilaterally contaminated normal model without nuisance parameter in its ability to fit real data sets, particularly in the presence of multimodality. However, as we have seen, this flexibility comes at the price of more complicated testing procedures. A simulation study can help us examine the performance of the testing procedure,

when BCN+NP model is correctly specified or an approximation to reality unless the omnibus null is true.

3.4 Simulation study

To verify empirically that our testing procedure for the omnibus null is consistent, i.e., the power of this test converges to 1 under H_a, a simulation study is done. This simulation study will also show the power for finite samples.

We take number of repeated samples n=100 and 1000, degree of freedom to be 5, 10, 50, 100, respectively. $\mu_1,\ \mu_2 \in \{-2,-1,0,1,2\}$, $\gamma_1,\ \gamma_2 \in \{0.1, 0.2\}$ and $(\alpha_1,\alpha_2,\alpha_3,\alpha_4) = (0.0125,0.0125,0.0125,0.0125), (0.005,0.02,0.02,0.005), (0.02,0.005,0.005,0.02)$, respectively. At each of the combinations, 1000 data sets were generated. Partial results are given in Table 3.1 - Table 3.6. Table 3.1 - Table 3.3 are simulated under the case where the population follows normal model or BCN+NP model.

Table 3.4 - Table 3.6 are simulated under the case where the population follows simple T distribution or T mixture model with nuisance parameter, then the data are transformed into quantities with normal null distribution as in microarray data analysis. More specifically, we define the *bilaterally contaminated T model with nuisance parameter* (BCT+NP) by the pdf

$$(1-\gamma_1-\gamma_2)\sigma^{-1}f_\nu(t/\sigma) + \gamma_1\sigma^{-1}f_\nu((t-\mu_1)/\sigma) + \gamma_2\sigma^{-1}f_\nu((t-\mu_2)/\sigma),$$

where f_ν denotes the T pdf on ν degree of freedom and F_ν denotes the corresponding cdf. This reduces to $f_\nu(t/\sigma)$ when $\gamma_1|\mu_1| + \gamma_2|\mu_2| = 0$. Data arising from this model can be transformed by $Z := \Phi^{-1}F_\nu(T)$, where Φ is the standard normal cdf. The

transformed data are then analyzed as if they had arise from the BCN+NP model, which is an approximation to reality.

Table 3.1: Type I Error When Population Follows Normal Distribution (Omnibus Null) and Power When Population Follows BCN+NP Model (Omnibus Alternative)

$(\mu_1, \mu_2, \gamma_1, \gamma_2)$ [a]	n=100	n=1000
(0,0,0,0) [b]	0.038	0.038
(1, -1, 0.1, 0.1)	0.050	0.054
(1, -1, 0.2, 0.2)	0.021	0.033
(2, -2, 0.1, 0.1)	0.092	0.521
(2, -2, 0.2, 0.2)	0.015	0.046
(1, 0, 0.2, 0.1)	0.291	1.000
(0, -1, 0.2, 0.1)	0.091	0.780
(2, 0, 0.2, 0.1)	0.845	1.000
(0, -2, 0.2, 0.1)	0.408	1.000
(1, -1, 0.2, 0.1)	0.096	0.657
(1, -2, 0.1, 0.1)	0.170	0.979
(2, -2, 0.2, 0.1)	0.137	0.975
(2, -1, 0.2, 0.2)	0.256	1.000

[a] $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.0125$
[b] Type I Error

To illustrate use of Table 3.1, we consider two examples. For example, when n=100 there is approximately a 3.8% Type I error probability associated with the Union-Intersection test of the omnibus null hypothesis when $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ = (0.0125, 0.0125, 0.0125, 0.0125). When n=100 there is approximately 29.1% power against the specific alternative $(\mu_1, \mu_2, \gamma_1, \gamma_2)$=(1, 0, 0.2, 0.1).

Table 3.2: Type I Error When Population Follows Normal Distribution (Omnibus Null) and Power When Population Follows BCN+NP Model (Omnibus Alternative)

$(\mu_1, \mu_2, \gamma_1, \gamma_2)$ [a]	n=100	n=1000
(0,0,0,0) [b]	0.037	0.046
(1, -1, 0.1, 0.1)	0.038	0.054
(1, -1, 0.2, 0.2)	0.037	0.042
(2, -2, 0.1, 0.1)	0.079	0.384
(2, -2, 0.2, 0.2)	0.029	0.039
(1, 0, 0.2, 0.1)	0.353	1.000
(0, -1, 0.2, 0.1)	0.101	0.854
(2, 0, 0.2, 0.1)	0.902	1.000
(0, -2, 0.2, 0.1)	0.443	1.000
(1, -1, 0.2, 0.1)	0.102	0.720
(1, -2, 0.1, 0.1)	0.182	0.988
(2, -2, 0.2, 0.1)	0.191	0.978
(2, -1, 0.2, 0.2)	0.275	1.000

[a] $\alpha_1 = \alpha_4 = 0.02, \alpha_2 = \alpha_3 = 0.005$
[b] Type I Error

To illustrate use of Table 3.2, we consider two examples. For example, when n=100 there is approximately a 3.7% Type I error probability associated with the Union-Intersection test of the omnibus null hypothesis when $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ = (0.02, 0.005, 0.005, 0.02). When n=100 there is approximately 35.3% power against the specific alternative $(\mu_1, \mu_2, \gamma_1, \gamma_2)$=(1, 0, 0.2, 0.1).

Table 3.3: Type I Error When Population Follows Normal Distribution (Omnibus Null) and Power When Population Follows BCN+NP Model (Omnibus Alternative)

$(\mu_1, \mu_2, \gamma_1, \gamma_2)$ [a]	n=100	n=1000
(0,0,0,0) [b]	0.032	0.038
(1, -1, 0.1, 0.1)	0.033	0.056
(1, -1, 0.2, 0.2)	0.024	0.024
(2, -2, 0.1, 0.1)	0.080	0.507
(2, -2, 0.2, 0.2)	0.002	0.055
(1, 0, 0.2, 0.1)	0.197	1.000
(0, -1, 0.2, 0.1)	0.073	0.661
(2, 0, 0.2, 0.1)	0.730	1.000
(0, -2, 0.2, 0.1)	0.319	1.000
(1, -1, 0.2, 0.1)	0.045	0.541
(1, -2, 0.1, 0.1)	0.138	0.958
(2, -2, 0.2, 0.1)	0.088	0.946
(2, -1, 0.2, 0.2)	0.135	0.999

[a] $\alpha_1 = \alpha_4 = 0.005, \alpha_2 = \alpha_3 = 0.02$
[b] Type I Error

To illustrate use of Table 3.3, we consider two examples. For example, when n=100 there is approximately a 3.2% Type I error probability associated with the Union-Intersection test of the omnibus null hypothesis when $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ = (0.005, 0.02, 0.02, 0.005). When n=100 there is approximately 19.7% power against the specific alternative $(\mu_1, \mu_2, \gamma_1, \gamma_2)$=(1, 0, 0.2, 0.1).

Table 3.4: Type I Error When Population Follows T Distribution and Power When Population Follows BCT+NP Model with Data Transformed to have Normal Null Distribution

	n=100				n=1000			
$(\mu_1, \mu_2, \gamma_1, \gamma_2)^a$	df=5	10	50	100	df=5	10	50	100
(0,0,0,0) [b]	0.027	0.030	0.026	0.039	0.038	0.037	0.045	0.051
(1, -1, 0.1, 0.1)	0.014	0.017	0.033	0.035	0.077	0.037	0.038	0.049
(1, -1, 0.2, 0.2)	0.022	0.012	0.016	0.019	0.528	0.236	0.030	0.024
(2, -2, 0.1, 0.1)	0.010	0.010	0.042	0.037	0.663	0.058	0.110	0.264
(2, -2, 0.2, 0.2)	0.014	0.013	0.015	0.010	1.000	1.000	0.319	0.132
(1, 0, 0.2, 0.1)	0.182	0.233	0.248	0.283	0.992	0.997	1.000	1.000
(0, -1, 0.2, 0.1)	0.060	0.078	0.088	0.091	0.560	0.592	0.725	0.740
(2, 0, 0.2, 0.1)	0.528	0.628	0.803	0.820	1.000	1.000	1.000	1.000
(0, -2, 0.2, 0.1)	0.143	0.197	0.341	0.403	0.979	0.996	1.000	1.000
(1, -1, 0.2, 0.1)	0.046	0.050	0.065	0.077	0.554	0.601	0.614	0.621
(1, -2, 0.1, 0.1)	0.026	0.049	0.103	0.140	0.539	0.686	0.934	0.955
(2, -2, 0.2, 0.1)	0.092	0.106	0.137	0.127	0.998	0.985	0.971	0.966
(2, -1, 0.2, 0.2)	0.089	0.096	0.172	0.203	0.999	0.998	0.999	1.000

[a] $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.0125$
[b] Type I Error

To illustrate use of Table 3.4, consider two examples. For example, when n=100 there is approximately 3.9% Type I error probability associated with applying the Union-Intersection test of the omnibus null hypothesis to transformed T statistics arising from an uncontaminated T distribution on 100 degrees of freedom when $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (0.0125, 0.0125, 0.0125, 0.0125)$. When n=100 there is approximately 28.3% power associated with applying the Union-Intersection test of the omnibus null hypothesis to transformed T statistics arising from the BCT+NP model

on 100 degrees of freedom when $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ = (0.0125, 0.0125, 0.0125, 0.0125) and $(\mu_1, \mu_2, \gamma_1, \gamma_2)$=(1, 0, 0.2, 0.1).

Table 3.5: Type I Error When Population Follows T Distribution and Power When Population Follows BCT+NP Model with Data Transformed to have Normal Null Distribution

	n=100				n=1000			
$(\mu_1, \mu_2, \gamma_1, \gamma_2)^a$	df=5	10	50	100	df=5	10	50	100
(0,0,0,0) [b]	0.038	0.038	0.055	0.034	0.045	0.045	0.044	0.048
(1, -1, 0.1, 0.1)	0.028	0.023	0.042	0.038	0.045	0.038	0.032	0.039
(1, -1, 0.2, 0.2)	0.026	0.017	0.024	0.028	0.330	0.136	0.043	0.036
(2, -2, 0.1, 0.1)	0.030	0.028	0.039	0.056	0.447	0.037	0.086	0.207
(2, -2, 0.2, 0.2)	0.032	0.025	0.027	0.022	1.000	0.997	0.165	0.063
(1, 0, 0.2, 0.1)	0.250	0.289	0.321	0.343	0.990	0.999	1.000	1.000
(0, -1, 0.2, 0.1)	0.067	0.084	0.116	0.106	0.585	0.707	0.792	0.810
(2, 0, 0.2, 0.1)	0.592	0.734	0.845	0.864	1.000	1.000	1.000	1.000
(0, -2, 0.2, 0.1)	0.177	0.269	0.406	0.044	0.986	0.999	1.000	1.000
(1, -1, 0.2, 0.1)	0.049	0.059	0.096	0.106	0.588	0.629	0.669	0.687
(1, -2, 0.1, 0.1)	0.054	0.076	0.141	0.152	0.539	0.733	0.955	0.966
(2, -2, 0.2, 0.1)	0.136	0.151	0.173	0.183	0.999	0.978	0.987	0.980
(2, -1, 0.2, 0.2)	0.119	0.157	0.243	0.276	0.998	0.993	0.999	1.000

[a] $\alpha_1 = \alpha_4 = 0.02, \alpha_2 = \alpha_3 = 0.005$
[b] Type I Error

To illustrate use of Table 3.5, consider two examples. For example, when n=100 there is approximately 3.4% Type I error probability associated with applying the Union-Intersection test of the omnibus null hypothesis to transformed T statistics arising from an uncontaminated T distribution on 100 degrees of freedom when

$(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ = (0.02, 0.005, 0.005, 0.02). When n=100 there is approximately 34.3% power associated with applying the Union-Intersection test of the omnibus null hypothesis to transformed T statistics arising from the BCT+NP model on 100 degrees of freedom when $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ = (0.02, 0.005, 0.005, 0.02) and $(\mu_1, \mu_2, \gamma_1, \gamma_2)$=(1, 0, 0.2, 0.1).

Table 3.6: Type I Error When Population Follows T Distribution and Power When Population Follows BCT+NP Model with Data Transformed to have Normal Null Distribution

	n=100				n=1000			
$(\mu_1, \mu_2, \gamma_1, \gamma_2)^a$	df=5	10	50	100	df=5	10	50	100
(0,0,0,0) [b]	0.029	0.034	0.020	0.034	0.034	0.044	0.027	0.029
(1, -1, 0.1, 0.1)	0.011	0.020	0.028	0.035	0.116	0.046	0.019	0.040
(1, -1, 0.2, 0.2)	0.010	0.012	0.012	0.014	0.636	0.294	0.019	0.029
(2, -2, 0.1, 0.1)	0.003	0.002	0.034	0.041	0.745	0.084	0.122	0.291
(2, -2, 0.2, 0.2)	0.019	0.006	0.002	0.009	1.000	1.000	0.464	0.212
(1, 0, 0.2, 0.1)	0.110	0.141	0.157	0.186	0.979	0.992	0.998	1.000
(0, -1, 0.2, 0.1)	0.031	0.038	0.064	0.048	0.392	0.496	0.635	0.635
(2, 0, 0.2, 0.1)	0.426	0.519	0.649	0.669	1.000	1.000	1.000	1.000
(0, -2, 0.2, 0.1)	0.088	0.110	0.233	0.254	0.954	0.995	1.000	1.000
(1, -1, 0.2, 0.1)	0.025	0.032	0.038	0.053	0.524	0.491	0.486	0.517
(1, -2, 0.1, 0.1)	0.019	0.033	0.069	0.109	0.479	0.558	0.900	0.925
(2, -2, 0.2, 0.1)	0.048	0.056	0.074	0.079	1.000	0.983	0.925	0.944
(2, -1, 0.2, 0.2)	0.041	0.060	0.106	0.087	0.999	0.992	0.999	0.999

[a] $\alpha_1 = \alpha_4 = 0.005, \alpha_2 = \alpha_3 = 0.02$
[b] Type I Error

To illustrate use of Table 3.6, consider two examples. For example, when n=100

there is approximately 3.4% Type I error probability associated with applying the Union-Intersection test of the omnibus null hypothesis to transformed T statistics arising from an uncontaminated T distribution on 100 degrees of freedom when $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ = (0.005, 0.02, 0.02, 0.005). When n=100 there is approximately 18.6% power associated with applying the Union-Intersection test of the omnibus null hypothesis to transformed T statistics arising from the BCT+NP model on 100 degrees of freedom when $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ = (0.005, 0.02, 0.02, 0.005) and $(\mu_1, \mu_2, \gamma_1, \gamma_2)$=(1, 0, 0.2, 0.1).

As shown in Tables 3.1 - 3.3, under normal distribution, the Type I error probability is approximately 0.038 for both sample size 100 and 1000 when $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ = (0.0125, 0.0125, 0.0125, 0.0125). If $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ = (0.02, 0.005, 0.005, 0.02) or $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ = (0.005, 0.02, 0.02, 0.005), the Type I error probability when sample size is 1000 (4.6% and 3.8%) is closer to 5% than when sample size is 100 (3.7% and 3.2%). Overall, the power with sample size 1000 is larger than sample size 100. The power under unilateral normal contamination is close to 1 when the sample size is 1000. If the sample size is relatively small, say 100, the unilateral test behaves better when the mean of normal contamination is large (± 2) than when it's small (± 1) and when the weight of normal contamination is large (0.2) than when it's small (0.1). The power under asymmetric bilateral normal contamination is fairly close to 1 when sample size is 1000. Also, the test does better when the the means of normal contaminations are different in absolute value than when the weights of normal contaminations are different. Under the symmetric bilateral normal contamination, the power is relatively low except (52.1%, 38.4%, 50.7%) when sample size is 1000 and the distribution has large means and small proportions, i.e. $(\mu_1, \mu_2, \gamma_1, \gamma_2) = (2, -2, 0.1, 0.1)$, regardless of the weights of α's.

Under T distribution with transformation(Tables 3.4-3.6), the Type I error probabilities are closer to 5% when sample size is 1000, and never much above 5%. The power overall is larger for bigger sample size. When sample size is 1000, the power under unilateral T contamination is not as good when the mean is close to 0 and proportion is small ($\mu_1 = -1, \gamma_1 = 0.1$), otherwise the power is close to 1. When sample size is 100, the power under unilateral T contamination is also larger if mean or proportion is large, but overall the power is still not as good when sample size is small. In both cases, the power seems to increase as the degrees of freedom increase from 5 to 100 regardless of the weights of α's. Under asymmetric bilateral T contamination with transformation, the power also increases as the degrees of freedom increase, but is smaller if either both means are close to 0 or both proportions are small. When sample size is 100, the power is fairly small, especially when $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (0.005,$ 0.02, 0.02, 0.005). The test with $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (0.02,$ 0.005, 0.005, 0.02) does the best of the three. When sample size is 1000, the power is larger but still not as good if the means are both close to 0 or the proportions are both small (0.1). When the means are both away from 0 or the proportions are both large (0.2), the power is close to 1. Under the symmetric bilateral T contamination, the power has no clear pattern but is fairly small when sample size is 100. When sample size is 1000, the power seems to decrease as degrees of freedom increase. Also, if both means are close to 0 and proportions are small, the power is much smaller than in the other cases. The test with $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (0.005,$ 0.02, 0.02, 0.005) does the best and the test with $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (0.02,$ 0.005, 0.005, 0.02) does the worst.

3.5 Application to Down's Syndrome Microarray Data Set

Down's syndrome is a genetic disorder caused by the presence of an extra piece of chromosome 21[44]. Down's syndrome is widely studied and is used to illustrate methods

for genetic data analysis[44]. To investigate the performance of BCT+NP model, we apply it to the Down's syndrome data set which is available via http://www.partek.com/Tutorials.

We take the gene information from four samples of cerebral tissue corresponding to Down's syndrome patients and seven samples corresponding to four control subjects. Linear mixed model is applied to the genes of patients and controls and T statistics are derived. Then the T statistics are transformed to Z statistics by the method mentioned in *Section 1.3*. To find a viable solution to the parameter estimating equations using method of moments, we used `optim` function in `R`. We try to minimize the quantity $f_n = \sum_{i=1}^{5}(g_i(\hat{\gamma}_1, \hat{\mu}_1, \hat{\gamma}_2, \hat{\mu}_2, \hat{\sigma}^2) - \hat{m}_i)^2$. If we can optimize this function such that its value is close to zero with no condition such as numerical convergence is violated, we have found a viable solution to the parameter estimating equations.

Among several trials on the complete data set using different starting values for the parameter estimates, the best result we got is $f_n = 1.67 \times 10^{-8}$ which is really close to zero. 1000 combinations of starting values are created. The starting value for μ_1 is randomly generated from $Unif(0, 3)$ and for μ_2 is randomly generated from $Unif(-3, 0)$, while γ_1 and γ_2 are generated from $Unif(0, 1)$ subject to $\gamma_1 + \gamma_2 < 1$. The lower bound for $(\mu_1, \gamma_1, \mu_2, \gamma_2, \sigma^2)$ is (0,0,-3,0,.1) while the upper bound is (3,1,0,1,1). `R` function `optim` is used and the optimizing method is *L-Broyden-Fletcher-Goldfarb-Shanno-B* which handles simple box constraints on variables. The final parameter estimates for $(\mu_1, \gamma_1, \mu_2, \gamma_2, \sigma^2)$, corresponding to the $f_n = 1.67 \times 10^{-8}$ identified above, are (1.596, 0.148, -1.530, 0.479,0.232).

Table 3.7 illustrates the optimal results for each individual chromosome. The same starting values, lower, upper bounds and optimizing method are used. As shown in the table, most chromosomes are well optimized in the sense of minimizing the

non-negative function f_n around 1.00E-08. Only Chromosome 13, 21 and Y have f_n above 0.0001, and among them Chromosome 21 has largest f_n value. Estimates of μ_1 and μ_2 are around 1.4 and -1.4 with estiamtes of γ_1 around 0.18 and γ_2 around 0.5 for most chromosomes. For Chromosome 21, both μ_1 and μ_2 estimates are extreme compared to other chromosomes. Estimate for μ_1 is 2.256 with estimated $\gamma_1 = 0.382$ and estimated μ_2 is -1.134 with estimated $\gamma_2 = 0.517$, which means Chromosome 21 has a moderate weight on a large mean on the over-expression side and a heavy weight on the under-expression side. This indicates Chromosome 21 has unusually pronounced over-expression which maybe attributable to the relationship of Chromosome 21 with the Down's syndrome.

The p-values for the Omnibus test of $N(0, \sigma^2)$ against UCN+NP/BCN+NP are all very small. This is because a simple normal model barely fits the data. The p-values for UCN+NP against BCN+NP are generated by first finding the value delta for each chromosome via simulation study, then plug in the delta values into an `R` function to obtain p-values. For detailed `R` code, refer to **Appendix A3**. The p-values are mostly significant at 0.01. Only Chromosomes 19 and 22 do not have p-values that are significant. Notice that Chromosome 19 has a relatively small proportion of over-expression and large proportion of under-expression. This makes it harder to detect the over-expression. As for Chromosome 22, the estimated μ_1 is relatively small with a moderate proportion, this also makes over-expression harder to detect. Chromosome Y is sort of an exception, the f_n function is not well optimized, so the parameter estimates may not be accurate. This may be due to the small sample size (41 valid genes) of genes on Chromosome Y. The parameter estimates of unknown category behave strangely, this may be because the f_n function is very flat around the maximizer. Since the unknown category contains genes from a variety of chromosomes, it's not readily interpretable.

Table 3.7: Parameter Estimates and P-values for Each Individual Chromosome

Chromosome	f_n [1]	$\hat{\mu}_1$	$\hat{\gamma}_1$	$\hat{\mu}_2$	$\hat{\gamma}_2$	$\hat{\sigma}^2$	p-value [2]	p-value [3]
unknown	7.75E-08	1.411	0.035	-1.412	0.831	0.282	< 0.0001	0.0002
1	3.10E-08	1.563	0.154	-1.548	0.466	0.221	< 0.0001	< 0.0001
2	1.63E-08	1.534	0.201	-1.533	0.435	0.229	< 0.0001	< 0.0001
3	2.33E-07	1.722	0.177	-1.640	0.392	0.107	< 0.0001	0.0011
4	1.39E-08	1.384	0.222	-1.403	0.472	0.351	< 0.0001	0.0006
5	3.66E-09	1.583	0.181	-1.526	0.410	0.229	< 0.0001	0.0016
6	7.37E-08	1.453	0.155	-1.439	0.546	0.272	< 0.0001	0.0001
7	3.53E-08	1.260	0.209	-1.403	0.506	0.347	< 0.0001	< 0.0001
8	4.18E-09	1.553	0.192	-1.506	0.429	0.250	< 0.0001	0.0033
9	1.03E-09	1.136	0.232	-1.269	0.554	0.400	< 0.0001	< 0.0001
10	5.80E-09	1.584	0.168	-1.591	0.405	0.216	< 0.0001	< 0.0001
11	8.33E-09	1.212	0.203	-1.370	0.549	0.353	< 0.0001	< 0.0001
12	3.19E-08	1.380	0.179	-1.422	0.503	0.268	< 0.0001	< 0.0001
13	6.75E-04	0.978	0.379	-1.185	0.521	0.489	0.0004	0.0013
14	1.23E-09	1.415	0.198	-1.541	0.474	0.230	< 0.0001	< 0.0001
15	3.50E-09	1.568	0.200	-1.517	0.438	0.225	< 0.0001	0.0065
16	2.01E-08	1.435	0.156	-1.485	0.493	0.248	< 0.0001	< 0.0001
17	6.59E-08	1.487	0.140	-1.497	0.520	0.257	< 0.0001	< 0.0001
18	1.25E-08	1.472	0.189	-1.493	0.397	0.229	< 0.0001	0.0029
19	4.47E-08	1.312	0.126	-1.444	0.601	0.294	< 0.0001	0.1992
20	3.58E-08	1.662	0.144	-1.600	0.401	0.169	< 0.0001	0.0001
21	3.88E-03	2.256	0.382	-1.134	0.517	0.557	0.0039	< 0.0001
22	6.20E-05	0.873	0.224	-1.308	0.676	0.422	< 0.0001	0.1434
X	9.61E-09	1.255	0.212	-1.436	0.535	0.332	< 0.0001	< 0.0001
Y	1.91E-03	0.000	0.041	-1.442	0.743	0.162	< 0.0001	0.0085

1. $f_n = \sum_{i=1}^{5}(g_i(\hat{\gamma}_1, \hat{\mu}_1, \hat{\gamma}_2, \hat{\mu}_2, \hat{\sigma}^2) - \hat{m}_i)^2$. 2. P-values for Omnibus test. 3. P-values for Unilateral test.

Chapter 4 Hierarchical Normal Mixture Model

4.1 Introduction

Hierarchical models (also called multilevel models) can be used for many fields of study. In sociology and biology, for example, hierarchical models can describe many different associations among variables at different levels through a hierarchical structure (see Snijders and Bosker 1999, Raudenbush and Bryk 2002). A common case in sociological studies based on demographic surveys is that the populations are usually clustered by geographical areas. The correlation among observations due to clustering can be modeled mathematically using multiple levels of the hierarchy. In such a scenario, a single level model cannot account for the effect of clustering and possibly is not appropriate.

In this Chapter, the idea of hierarchical modeling is incorporated into a normal mixture model. This model is somewhat different from the BCN+NP model considered earlier and has the following structure:

$$Y|X = x, M = m \;\sim\; N(a_m + b_m x, \tau^2)$$

$$X|M = m \;\sim\; N(\mu_m, \sigma^2)$$

where a_m, b_m are intercept and slope of regressed mean and τ^2 is the variance at the higher level, given that $M = m$. Note that M is the categorical variable identifying which mixture component (X, Y) belongs to. μ_m is the mean of X in the m^{th} component and σ^2 is the variance, presumed common across components. I call this kind of multilevel model the *hierarchical normal mixture model with nuisance parameters* (HNM+NP).

As noted above, the HNM+NP model is rather different from the BCN+NP model, not only because of the multiple levels but also because the number of components is not limited to three (and, indeed, may be unknown a priori). Moreover, we no longer assume that there is a primary component whose mean is known.

This chapter will describe how to approximate the maximum likelihood estimates of parameters in HNM+NP model through EM-algorithm. New criteria other than AIC, BIC for choosing the number of components will be introduced, followed by some simulation study. Finally the model will be applied to a birthweight data set to study the relation between children's birthweight and mortality.

4.2 Parameter Estimation Using the EM Algorithm

The EM algorithm is widely used for approximating maximum likelihood estimates of parameters from a mixture model (see Dempster, Laird and Rubin 1977, Redner and Walker 1984, Bilmes 1998). For the HNM+NP model, we can also apply the EM algorithm to find parameter estimates. Let $i \in \{1, 2, ..., n\}$ be the indices for subjects, K be the presumed total number of components, $\pi_m \geq 0$ be the probability of a subject belonging to component m with the constraint that $\sum_{m=1}^{K} \pi_m = 1$ and M_i be the indicator for which component subject i belongs to. Let $\boldsymbol{\theta}_m = (\mu_m, \sigma^2, a_m, b_m, \tau^2)$ denote the vector of m^{th} component parameters that we want to estimate. Then the contribution to the "complete data" likelihood function from component m is

$$L(\boldsymbol{\theta}_m|\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{x}, a_m, b_m, \tau^2) f(\boldsymbol{x}|\mu_m, \sigma^2)$$
$$= \prod_{i=1}^{n} \left\{ \mathbb{I}_{\{M_i=m\}} \frac{\pi_m}{2\pi\sigma\tau} \exp\left[-\frac{1}{2} \left(\frac{(x_i - \mu_m)^2}{\sigma^2} + \frac{(y_i - a_m - b_m x_i)^2}{\tau^2} \right) \right] \right\}$$

where M_i is not observed in practice. So we replace $\mathbb{I}_{\{M_i=m\}}$ by its conditional expectation.

Using Bayes' Theorem, we find that

$$\boldsymbol{E}[\mathbb{I}_{\{M_i=m\}}|x_i,y_i] = \boldsymbol{P}[M_i=m|x_i,y_i] \tag{4.1}$$

$$= \frac{f_{X,Y|M}(x_i,y_i|M_i=m)\boldsymbol{P}[M_i=m]}{\sum_{m'=1}^{K} f_{X,Y|M}(x_i,y_i|M_i=m')\boldsymbol{P}[M_i=m']} \tag{4.2}$$

$$= \frac{\pi_m f_{X,Y|\boldsymbol{\theta}_m}(x_i,y_i|\boldsymbol{\theta}_m)}{\sum_{m'=1}^{K} \pi_{m'} f_{X,Y|\boldsymbol{\theta}_{m'}}(x_i,y_i|\boldsymbol{\theta}_{m'})}. \tag{4.3}$$

Define $w_{im} := \boldsymbol{E}[\mathbb{I}_{\{M_i=m\}}|x_i,y_i]$, then the complete data log likelihood function is approximated by

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{m=1}^{K}\left\{w_{im} * \log\left(\frac{\pi_m}{2\pi\sigma\tau}\exp\left[-\frac{1}{2}\left(\frac{(x_i-\mu_m)^2}{\sigma^2}+\frac{(y_i-a_m-b_m x_i)^2}{\tau^2}\right)\right]\right)\right\} \tag{4.4}$$

$$= \sum_{i=1}^{n}\sum_{m=1}^{K}\left\{w_{im}\left[\log(\pi_m)-\log(2\pi\sigma\tau)-\frac{1}{2}\left(\frac{(x_i-\mu_m)^2}{\sigma^2}+\frac{(y_i-a_m-b_m x_i)^2}{\tau^2}\right)\right]\right\} \tag{4.5}$$

$$= \sum_{i=1}^{n}\sum_{m=1}^{K}\left\{w_{im}\left[\log(\pi_m)-\log(\sigma\tau)-\frac{1}{2}\left(\frac{(x_i-\mu_m)^2}{\sigma^2}+\frac{(y_i-a_m-b_m x_i)^2}{\tau^2}\right)\right]\right\} + C, \tag{4.6}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K)$ and C is free of $\boldsymbol{\theta}$.

Then the function

$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) := \boldsymbol{E}[l(\boldsymbol{\theta})|\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}^{(t)}]$

$$= \sum_{i=1}^{n}\sum_{m=1}^{K} w_{im}^{(t)}\left[\log(\pi_m)-\log(\sigma\tau)-\frac{1}{2}\left(\frac{(x_i-\mu_m)^2}{\sigma^2}+\frac{(y_i-a_m-b_m x_i)^2}{\tau^2}\right)\right] + C$$

represents the evaluation of the aforementioned approximation when w_{im} is evaluated at $\boldsymbol{\theta}^{(t)}$. We denote this by $w_{im}^{(t)}$ with the interpretation that $\boldsymbol{\theta}^{(t)}$ represents the latest estimate of $\boldsymbol{\theta}$ after t iterations of the EM algorithm.

After differentiation with respect to each component of $\boldsymbol{\theta}$, the updating equations to produce new estimates of $\boldsymbol{\theta}$ after $t+1$ iterations are

$$\pi_m^{(t+1)} = \frac{\sum_i w_{im}^{(t)}}{n}$$

$$\mu_m^{(t+1)} = \frac{\sum_i w_{im}^{(t)} x_i}{\sum_i w_{im}^{(t)}}$$

$$\sigma_m^{(t+1)} = \sqrt{\frac{\sum_i w_{im}^{(t)} (x_i - \mu_m^{(t+1)})^2}{\sum_i w_{im}^{(t)}}}$$

$$\tau_m^{(t+1)} = \sqrt{\frac{\sum_i w_{im}^{(t)} (y_i - a_m^{(t+1)} - b_m^{(t+1)} x_i)^2}{\sum_i w_{im}^{(t)}}}$$

$$\begin{bmatrix} a_m^{(t+1)} \\ b_m^{(t+1)} \end{bmatrix} = \begin{bmatrix} \sum_i w_{im}^{(t)} & \sum_i w_{im}^{(t)} x_i \\ \sum_i w_{im}^{(t)} x_i & \sum_i w_{im}^{(t)} x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_i w_{im}^{(t)} y_i \\ \sum_i w_{im}^{(t)} x_i y_i \end{bmatrix}$$

Let $v_1 = \frac{1}{n}\sum_i w_{i1}^{(t)}$, $v_2 = \frac{1}{n}\sum_i w_{i2}^{(t)}$, etc. Then the pooled variance

$$\begin{aligned}
(\hat{\sigma}^2)^{(t+1)} &= \sum_m v_m (\hat{\sigma}_m^2)^{(t+1)} \\
&= \frac{1}{n} \sum_m \frac{\left(\sum_i w_{im}^{(t)}\right) \sum_i w_{im}^{(t)} (x_i - \mu_m^{(t+1)})^2}{\sum_i w_{im}^{(t)}} \\
&= \frac{1}{n} \sum_{i=1}^{n} \sum_{m=1}^{K} w_{im}^{(t)} (x_i - \mu_m^{(t+1)})^2
\end{aligned}$$

Similarly,

$$(\hat{\tau}^2)^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \sum_{m=1}^{K} w_{im}^{(t)} (y_i - a_m^{(t+1)} - b_m^{(t+1)} x_i)^2$$

The correlation between X and Y can be obtained as follows,

$$\begin{aligned}
E[XY|M=m] &= \int_R \int_R xy f(x,y|m) dx dy \cdot \mathrm{I}_{(M=m)} \\
&= \int_R \frac{x}{\sqrt{2\pi}\sigma_m} \exp\left(-\frac{(x-\mu_m)^2}{2\sigma_m^2}\right) \int_R \frac{y}{\sqrt{2\pi}\tau_m} \exp\left(-\frac{(y-a_m-b_m x)^2}{2\tau_m^2}\right) dy dx \cdot \mathrm{I}_{(M=m} \\
&= \int_R \frac{x}{\sqrt{2\pi}\sigma_m} \exp\left(-\frac{(x-\mu_m)^2}{2\sigma_m^2}\right) (a_m + b_m x) dx \cdot \mathrm{I}_{(M=m)} \\
&= \int_R \frac{x a_m}{\sqrt{2\pi}\sigma_m} \exp\left(-\frac{(x-\mu_m)^2}{2\sigma_m^2}\right) dx \cdot \mathrm{I}_{(M=m)} \\
&\quad + \int_R \frac{x^2 b_m}{\sqrt{2\pi}\sigma_m} \exp\left(-\frac{(x-\mu_m)^2}{2\sigma_m^2}\right) dx \cdot \mathrm{I}_{(M=m)} \\
&= a_m \mu_m + b_m (\sigma_m^2 + \mu_m^2)
\end{aligned}$$

$$\rho_{X,Y|M} = \text{Corr}(X, Y|M = m) = \frac{E[(X - \mu_X)(Y - \mu_Y)|M = m]}{\sigma_{X|M}\sigma_{Y|M}}$$
$$= \frac{E[XY|M = m] - \mu_{X|M}\mu_{Y|M}}{\sigma_{X|M}\sigma_{Y|M}}$$
$$= \frac{a_m\mu_m + b_m(\sigma_m^2 + \mu_m^2) - \mu_m(a_m + b_m\mu_m)}{\sigma_m\sigma_{Y|M}}$$
$$= \frac{b_m\sigma_m}{\sigma_{Y|M}}$$

$$\sigma_{Y|M} = Var[Y|M] = E[Var[Y|X, M]] + Var[E[Y|X, M]]$$
$$= \tau_m^2 + Var[a_m + b_mX|M]$$
$$= \tau_m^2 + b_m^2 Var[X|M]$$
$$= \tau_m^2 + b_m^2\sigma_m^2$$

$$\rho_{X,Y|M} = \frac{b_m\sigma_m}{\sqrt{\tau_m^2 + b_m^2\sigma_m^2}}$$

4.3 Singular Bayesian Information Criterion

Many papers in the statistical literature discussed methods of model selection. Two well-known approaches are the information-theoretic selection based on Kullback-leibler (K-L) information loss and Bayesian model selection based on Bayes factors (Burnham, Anderson 2004). Akaike Information Criteria (AIC) represents the first approach and Bayesian Information Criteria (BIC) of Schwarz (1978) represents the second approach.

However, due to the singularity problem in mixture models, neither AIC nor BIC may be an appropriate criterion. AIC doesn't put enough penalty on the number of parameters, thereby tending to overestimate the number of mixing components. On the other hand, the determinants of Fisher-information matrices for mixture models

may be singular and thus non-invertible.

Although BIC is known to be consistent in some singular settings, the technical arguments in its Bayesian-inspired derivation do not apply (Keribin 2000, Drton and Martyn 2013).

Drton proposed a new information criterion for singular models called the *singular Bayesian information criterion* (sBIC). This criterion conquers the difficulty that the determinants of Fisher-information matrices for mixture models may be singular. Drton also showed that sBIC is consistent.

Suppose $Y_n = (Y_{n1}, ..., Y_{nn})$ is a sample of n iid observations and $\{M_i, i \in I\}$ is a finite set of candidate models. For each model M_i, $P(M_i)$ is the positive prior probability and $P(\pi_i|M_i)$ is prior distribution for data-generating distribution $\pi_i \in M_i$. $P(Y_n|\pi_i, M_i)$ is the likelihood of Y_n under data-generating distribution π_i from model M_i. The marginal likelihood of M_i is

$$L(M_i) := P(Y_n|M_i) = \int_{M_i} P(Y_n|\pi_i, M_i) dP(\pi_i|M_i).$$

The posterior model probability is

$$P(M_i|Y_n) \propto P(M_i)L(M_i).$$

By Schwarz's theorem (Schwarz 1978), the Bayesian information criterion for model M_i is

$$\mathrm{BIC}(M_i) = \log P(Y_n|\hat{\pi}_i, M_i) - \frac{d_i}{2}\log(n),$$

where $\hat{\pi}_i$ is the data-generating distribution as estimated by maximum likelihood. However, this is generally not valid in singular models.

Before proceeding to the derivation of sBIC, it is necessary to introduce the learning coefficient and multiplicity. Suppose $K(w)$ is the K-L divergence,

$$K(w) = \int f(x) \log \frac{f(x)}{g(x|w)} dx,$$

where $f(x)$ is the true probability density function and $g(x|w)$ is a learning machine. A learning machine refers to the statistical model of interest. Then the zeta function is defined as

$$\zeta(z) = \int K(w)^z p(w) dw,$$

where $p(w)$ is an *a priori* probability density function. Watanabe (2009) states that $-\lambda$ and m are the largest pole of the zeta function $\zeta(z)$ and its order, then the positive λ is called a learning coefficient and m is called a multiplicity. In Conway's 1973 book, he defined pole and its order. If $z = a$ is an isolated singularity of f, then a is a *pole* of f if $\lim_{z \to a} |f(z)| = \infty$. If f has a pole at $z = a$ and m is the smallest positive integer such that $f(z)(z-a)^m$ has a removable singularity at $z = a$, then f has a *pole of order* m at $z = a$.

Watanabe (2009 Theorem 6.7) proved that for most singular models,

$$\log L(M_i) = \log P(Y_n|\pi_0, M_i) - \lambda_i(\pi_0) \log(n) + [m_i(\pi_0) - 1] \log\log(n) + Op(1),$$

where π_0 is the true data-generating distribution, $\lambda_i(\pi_0)$ is the *learning coefficient* and $m_i(\pi_0)$ is its *multiplicity*.

If the sequence of likelihood ratios $P(Y_n|\hat{\pi}_i, M_i)/P(Y_n|\pi_0, M_i)$ is bounded in probability, then we also have that

$$\log L(M_i) = \log P(Y_n|\hat{\pi}_i, M_i) - \lambda_i(\pi_0) \log(n) + [m_i(\pi_0) - 1] \log\log(n) + Op(1).$$

where $\lambda_i(\pi_0) \in (0, d_i/2]$ is learning coefficient and $m_i(\pi_0) \in \{1, ..., d_i\}$ is multiplicity.

Drton (2009) showed that the likelihood ratios for singular submodels of exponential families will converge in distribution and are thus bounded in probability. Azaïs *et al* (2006 2009) proved that for more complicated models like mixture models, likelihood ratios converge in distribution under compactness assumptions on the parameter space.

Let M_i be the candidate model, then the singular Bayesian information criterion for model M_i is

$$\text{sBIC}(M_i) = \log L'(M_i),$$

where $L'(M_i) : i \in I$ is the unique solution to an equation system which involves the learning coefficients of singular models. Watanabe also proved if a model maximizes sBIC, then the probability that this model is a true model of minimal Bayes complexity (and thus also a smallest true model) tends to 1 as $n \to \infty$ under some mild assumptions.

If true data-generating distribution π_0 is known, the marginal likelihood becomes

$$L'_{\pi_0}(M_i) := P(Y_n|\hat{\pi}_i, M_i)n^{-\lambda_i(\pi_0)}(\log n)^{m_i(\pi_0)-1}.$$

However, if the data-generating distribution is unknown, Drton (2013) proposes to assign a probability distribution Q_i to the distributions in model M_i. He then eliminates the unknown distribution π_0 by marginalization and computes the approximated marginal likelihood

$$L'_{Q_i}(M_i) := \int_{M_i} L'_{\pi_0}(M_i) dQ_i(\pi_0).$$

In regular case,

$$L'_{Q_i}(M_i) = e^{\text{BIC}(M_i)}$$

with $\lambda_i = d_i/2, m_i = 1$ for all probability measures Q_i on M_i. In singular case, Drton advocates the use of the posterior distribution

$$Q_i(\pi_0) := P(\pi_0|\{M : M \subseteq M_i\}, Y_n) = \frac{\sum_{j \preceq i} P(\pi_0|M_j, Y_n)P(M_j|Y_n)}{\sum_{j \preceq i} P(M_j|Y_n)}$$

obtained by conditioning on the family of all submodels of M_i.

Suppose the random probability measure π_0 in $M_j \subseteq M_i$ is distributed according to the posterior distribution $P(\pi_0|M_i, Y_n)$. Then it holds that both $\lambda_i(\pi_0)$ and $m_i(\pi_0)$ are almost surely constant under some conditions. For $j \preceq i$, let λ_{ij} and m_{ij} denote these constants and define

$$L'_{ij} := P(Y_n|\hat{\pi}_i, M_i)n^{-\lambda_{ij}}(\log n)^{m_{ij}-1} > 0,$$

which can be evaluated in statistical practice. Let $L'(M_i) := L'_{Q_i}(M_i)$ when Q_i is chosen as $Q_i(\pi_0)$, we have

$$L'(M_i) = \frac{\sum_{j \preceq i} L'_{ij} L(M_j)P(M_j)}{\sum_{j \preceq i} L(M_j)P(M_j)}$$

Replacing $L(M_j)$ by $L'(M_j)$,

$$L''(M_i) = \frac{\sum_{j \preceq i} L'_{ij} L'(M_j)P(M_j)}{\sum_{j \preceq i} L'(M_j)P(M_j)}$$

where $L''(M_i)$ is just another notation for approximated marginal likelihood to avoid confusion. Drton (2013) showed the equation system above has a unique solution with all unknowns $L''(M_i) > 0$. If i is a minimal element of I, then $j \preceq i$ implies $j = i$ and the equation has the unique positive solution

$$L''(M_i) = L'_{ii} > 0,$$

which coincides with the exponential of the usual BIC for model M_i.

Otherwise, if i is not a minimal element of I and thus there exists $j \prec i$, then

$$L'(M_i) = \frac{1}{2}\left(-b_i + \sqrt{b_i^2 + 4c_i}\right)$$

with

$$b_i = -L'_{ii} + \sum_{j \prec i} L'(M_j)\frac{P(M_j)}{P(M_i)},$$

$$c_i = \sum_{j \prec i} L'_{ij} L'(M_j)\frac{P(M_j)}{P(M_i)}.$$

Hence, for $i \in I$, $L'(M_i)$ can be derived recursively. Then the singular Bayesian information criterion for model M_i is

$$\text{sBIC}(M_i) = \log L'(M_i) = \log P(Y_n|\hat{\pi}_i, M_i) - \text{penalty}(M_i),$$

where $\text{penalty}(M_i) \leq \dim(M_i)/2 \cdot \log(n)$. Hence, $\text{penalty}(M_i)$ is milder than that of BIC.

To calculate sBIC for HNM+NP model, we need to find the learning coefficient and multiplicity. Suppose i is the number of normal mixture components in the learning machine (Watanabe 2009) and j is the number of normal mixture components of the true distribution. By Watanabe (2009 Section 7.3), the dimension of the parameter space is $2i$ for the lower level (the counts for μ_m, γ_m and σ^2 are $i, i-1, 1$) and for the higher level is $2i+1$ (the counts for a_m, b_m and τ^2 are $i, i, 1$). The number of free parameters in total is $3(i-j)$ (the free parameters come from a_m, b_m, γ_m). The degree of freedom is the difference of total dimension of parameter space minus number of free parameters, which is $3j + i + 1$. Hence we have

$$\lambda_{ij} \leq 0.5(3j + i + 1)$$

by Watanabe.

4.4 Consistency of sBIC

In Drton's paper (2013), he proved the consistency of sBIC under some assumptions. In order to show that sBIC is consistent when applied to HNM+NP model, we need to verify that HNM+NP model meets the following assumptions:

1. The sequence of likelihood ratios of any two true models is bounded in probability as $n \to \infty$,
2. For a true model M_i and a false model M_k, there exists a positive constant $\epsilon_{ik} > 0$, such that the probability of the sequence of likelihood ratios for false model vs true model is $\leq e^{-\epsilon_{ik} n}$ tends to 1 as $n \to \infty$,
3. For any two true models M_i and M_k and any of their corresponding true submodels M_j and M_l, the Bayes complexity should be monotonically increasing, i.e., $(\lambda_{ij}, m_{ij}) < (\lambda_{kl}, m_{kl})$ if $i \prec k$ and $j \prec l$.

There's another assumption which is not used for proving consistency but is useful to investigate the Bayesian behavior of sBIC. It says, for a true model M_i, its Bayes complexity should be nondecreasing in true submodels, i.e., if $j \preceq k$ index two true submodels of M_i, then $(\lambda_{ij}, m_{ij}) < (\lambda_{ik}, m_{ik})$.

Before showing HNM+NP model satisfies all aforementioned assumptions, it's necessary to state some definitions. First, a model M_i is a true model if the data-generating distribution $\pi_0 \in M_i$. Otherwise, M_i is called a false model. A true model is called the smallest true model if all its strict submodels are false. Moreover, when comparing the Bayes complexity of two true models, the operator '$<$' is the lexicographic order on $\mathbb{R}^2$. So $(a, b) \leq (c, d)$ means either $a < c$ or $a = c$ and $b \leq d$.

To show assumption 1, suppose M_f and M_g are two true models. M_h is the smallest

true model. Define the quantity

$$\lambda_{ab} = \log \frac{P(Y_n|\hat{\pi}_a, M_a)}{P(Y_n|\hat{\pi}_b, M_b)}.$$

Then Dacunha-Castelle and Gassiat (1999) showed that the likelihood ratio test statistic

$$\lambda_{fh} \xrightarrow{L} \frac{1}{2} \sup_{d \in \mathcal{D}} (\xi_d)^2 \cdot 1_{\xi_d \geq 0},$$

where ξ_d is a Gaussian process indexed by $d \in \mathcal{D}$, which is assumed compact. This implies λ_{fh} is also bounded in probability. Since f is just an arbitrary index for any true model, the above result is also true for M_g. That is,

$$\lambda_{gh} \xrightarrow{L} \frac{1}{2} \sup_{d \in \mathcal{D}} (\xi'_d)^2 \cdot 1_{\xi'_d \geq 0}.$$

Thus we have $\lambda_{fg} = \log \dfrac{P(Y_n|\hat{\pi}_f, M_f)}{P(Y_n|\hat{\pi}_g, M_g)}$ is also bounded in probability. □

Assumption 2 is more complex than assumption 1. Suppose M_f is a false model, M_t is a true model and M_h is the smallest true model. The quantity

$$\begin{aligned} \lambda_{fh} &= \log \frac{P(Y_n|\hat{\pi}_f, M_f)}{P(Y_n|\hat{\pi}_h, M_h)} \\ &= \sum_{i=1}^{n} \log f_{M_f}(Y_i|\hat{\pi}_f) - \sum_{i=1}^{n} \log f_{M_t}(Y_i|\hat{\pi}_t) \end{aligned}$$

Since $\hat{\pi_h} \xrightarrow{a.s.} \pi_h$ and there exists π_f such that $\hat{\pi_f} \xrightarrow{a.s.} \pi_f$, then assuming compact parameter spaces

$$\frac{1}{n} \sum_{i=1}^{n} \log f_{M_f}(Y_i|\hat{\pi}_f) \xrightarrow{a.s.} E \log f_{M_f}(Y_i|\pi_f),$$

by strong law of large number and likewise

$$\frac{1}{n} \sum_{i=1}^{n} \log f_{M_h}(Y_i|\hat{\pi}_h) \xrightarrow{a.s.} E \log f_{M_h}(Y_i|\pi_h).$$

Then by Slutsky's theorem and Jensen's inequality,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\log f_{M_f}(Y_i|\hat{\pi}_f) - \frac{1}{n}\sum_{i=1}^{n}\log f_{M_h}(Y_i|\hat{\pi}_h) &\xrightarrow{a.s.} E\log f_{M_f}(Y_i|\pi_f) - E\log f_{M_h}(Y_i|\pi_h)\\ &= \int \log\frac{f_{M_f}(Y_i|\pi_f)}{f_{M_h}(Y_i|\pi_h)} f_{M_h}(Y_i|\pi_h)dy\\ &< \log\int \frac{f_{M_f}(Y_i|\pi_f)}{f_{M_h}(Y_i|\pi_h)} f_{M_h}(Y_i|\pi_h)dy\\ &= \log(1) = 0\end{aligned}$$

Define δ by $\frac{1}{n}\sum_{i=1}^{n}\log f_{M_f}(Y_i|\hat{\pi}_f) - \frac{1}{n}\sum_{i=1}^{n}\log f_{M_h}(Y_i|\hat{\pi}_h) \xrightarrow{a.s.} -2\delta$, then for sufficiently large n (possibly depending on the underlying element of the probability space)

$$\sum_{i=1}^{n}\log f_{M_f}(Y_i|\hat{\pi}_f) - \sum_{i=1}^{n}\log f_{M_h}(Y_i|\hat{\pi}_h) \le -n\cdot\delta.$$

Taking exponential of both sides we have the probability of

$$\frac{P(Y_n|\hat{\pi}_f, M_f)}{P(Y_n|\hat{\pi}_h, M_h)} \le e^{-n\cdot\delta}$$

tends to 1 as $n \to \infty$.

Moreover, M_h is the smallest true model, we have the proportion

$$\frac{P(Y_n|\hat{\pi}_h, M_h)}{P(Y_n|\hat{\pi}_t, M_t)} \le 1,$$

Thus we have

$$\frac{P(Y_n|\hat{\pi}_f, M_f)}{P(Y_n|\hat{\pi}_t, M_t)} = \frac{P(Y_n|\hat{\pi}_f, M_f)}{P(Y_n|\hat{\pi}_h, M_h)}\cdot\frac{P(Y_n|\hat{\pi}_h, M_h)}{P(Y_n|\hat{\pi}_t, M_t)} \le e^{-n\cdot\delta}$$

□

Assumption 3 is pretty straight forward. For the HNM+NP model, the learning coefficient $\lambda_{ij} = \frac{1}{2}(3j+i+1)$ which is larger than λ_{kl} if $i \prec k$ and $j \preceq l$. □

By Drton (2013) Theorem 4.1, if assumption 1-3 are satisfied and suppose $M_{\hat{i}}$ is chosen by maximizing sBIC, i.e.,

$$\hat{i} = \arg\max_{i \in I} \text{sBIC}(M_i).$$

Then the probability that $M_{\hat{i}}$ is a true model of minimal Bayes complexity tends to 1 as $n \to \infty$ and thus sBIC applied to HNM+NP is consistent.

4.5 Simulation Study

We generate data that follow a two-level hierarchical normal mixture distribution. The lower level of the data X_i is generated from an M component normal mixture model with mean μ_m in component m and variance σ^2. The higher level of the data Y_i, given $X_i = x_i$ and membership in component m, is generated from a normal model with mean $a_m + b_m x_i$ and variance τ^2. We take sample sizes $n = 1000$, number of components $M \in \{2, 3, 4, 5\}$ and allow $M = 1$ just for comparison purpose. Other parameters are as shown in Table 4.1. For each combination of M and n, we generate 50 data sets and record the number of times that 2, 3, 4 and 5 component models were selected respectively by AIC, BIC and sBIC.

One thing worth mentioning is, due to the limit of `R` software, the algorithm for calculating sBIC showed in Section 4.3 is not practically feasible when sample size is large (e.g. $n \geq 500$). Thus we apply Lemma 4.1 in Drton' paper (2013) which states under assumption 2, if M_i is a smallest true model, then

$$\text{sBIC}(M_i) = \log(L'_{ii}) + o_p(1).$$

Applying this lemma, we have approximated $\text{sBIC}(M_i) \approx \log(L'_{ii})$ and avoided calculating the likelihood which is a product of n floating point numbers. This lemma is also used when fitting HNM+NP model to the Vital Statistics Natality Birth Data.

Table 4.1: Parameter Combinations of HNM+NP Model for Simulation Study

M	2	3	4	5	M	2	3	4	5
π_1	0.4	0.2	0.2	0.1	a_1	-2	-2	-2	-2
π_2	0.6	0.5	0.3	0.2	a_2	2	1	-1	-1
π_3	-	0.3	0.3	0.3	a_3	-	2	1	0
π_4	-	-	0.2	0.2	a_4	-	-	2	1
π_5	-	-	-	0.2	a_5	-	-	-	2
μ_1	-2	-2	-2	-2	b_1	-2	-2	-2	-2
μ_2	2	-1	-1	-1	b_2	2	1	-1	-1
μ_3	-	2	1	0	b_3	-	2	1	0
μ_4	-	-	3	1	b_4	-	-	2	1
μ_5	-	-	-	2	b_5	-	-	-	2
σ^2	1	1	1	1	τ^2	2	2	2	2

The results from simulation study are shown in Table 4.2. When the samples are generated from a two component normal mixture population (M=2), all three criteria agreed to choose two component normal mixture model correctly. When M=2, the two modes may be easy to distinguish and all the criteria can choose the right model. There exists some disagreement when M=3, around 40% of the samples choose three component model correctly, 60% of the samples choose two component model. None of the criteria do a good job when the population comes from a 4 or 5 component normal mixture. Since the modes could be very close to each other, it's even harder to detect them, not even in favor of the 3 component model.

Table 4.2: Number of Components Chosen by AIC, BIC and sBIC, out of 50 samples of size 1000

	M=2				M=3				M=4				M=5			
Criterion	2	3	4	5	2	3	4	5	2	3	4	5	2	3	4	5
AIC	50	0	0	0	30	20	0	0	50	0	0	0	50	0	0	0
BIC	50	0	0	0	31	19	0	0	50	0	0	0	50	0	0	0
sBIC	50	0	0	0	30	20	0	0	50	0	0	0	50	0	0	0

The unsatisfying results may be due to the modest differences between component means, also can be the inadequate sample size relative to model complexity. Our focus of the HNM+NP model is on how Y is linearly related to X, more specifically, is the mean of Y positively or negatively related to X within each component. We will see that in the next section where we applied HNM+NP model to the Birth Weight data set.

4.6 Application to Vital Statistics Natality Birth Data Set

In this section, the HNM+NP model will be applied to the NCHS's Vital Statistics Natality Birth Data from 2011. Natality Data from the National Vital Statistics System of the National Center for Health Statistics provide demographic and health data for births occurring during the indicated calendar year. The data can be downloaded from this link:
`http://www.nber.org/data/vital-statistics-natality-data.html`.

Our interest is to investigate the relationship between birth weight (in grams) of infants and the estimated obstetric gestation (in weeks). The birth weight ranges from 227 to 8165 grams. A 9999 indicates unknown birth weight. Meanwhile, the estimated obstetric gestation ranges from 17 to 47 weeks and a 99 means unknown

or not stated. The missing or unknown data are removed. We will randomly draw subsets of sizes 100, 500, 2500, 5000. We will fit 2,3,4 and 5 component HNM+NP models via EM algorithm and use AIC, BIC and sBIC to estimate the "true" number of components. Looking at the multiple sample sizes will allow us to assess sensitivity of the results to n, and examining multiple subsets at a given n will allow us to assess reproducibility. Considering the nature of the two variables, we will use estimated obstetric gestation to fit the lower level and birth weight to fit the higher level, since "predicting" birth weight from obstetric gestation seems more reasonable than "predicting" obstetric gestation from birth weight. Ten samples of each sample size are randomly selected and the number of components chosen by AIC, BIC and sBIC are listed in Table 4.3.

Table 4.3: Numbers of Samples that Choose 2-, 3-, 4-, 5-Component Models

	AIC				BIC				sBIC			
Sample Size	2	3	4	5	2	3	4	5	2	3	4	5
n=100	5	4	0	1	6	4	0	0	5	4	0	1
n=500	5	3	2	0	5	3	2	0	5	3	2	0
n=2500	3	1	5	1	3	1	5	1	3	1	5	1
n=5000	2	2	4	2	2	2	4	2	2	2	4	2

Table 4.3 consists of model selection results of 10 samples for each sample size, 40 samples in total. When sample size is in $\{100, 500\}$, AIC, BIC and SBIC all tend to suggest a 2-component hierarchical normal mixture model. This is a very common phenomenon in mixture modeling that a 2- or 3-component model is preferred when the sample size is not sufficiently large. The same thing happens here for hierarchical normal mixture model. When n is small, the data has less variability that is clearly ascribed to heterogeneity and is not as a good representation of the population as

large n when the population has complexity larger than 3. When the sample size rises to 2500 and 5000, the AIC, BIC and sBIC tend to choose a 4-component hierarchical normal mixture model more consistently. For the infant birth weight data, it makes more sense that the birth weights follow a 4-component HNM model than a 2-component HNM model. The 4-component model will group infants into more detailed gestational age and birth weight clusters and thus have a better description of the joint distribution of gestational age and birth weight.

Note that the results might be different if we allow heteroscedasticity (i.e. σ and τ vary by mixture component). Also, gestation ages and birthweights have been rounded, so that sample of size 100 displays digit preference.

We now present more detailed results from one additional generated random sample for n=100, 500, 2500, 5000. The parameter estimates for each sample size are listed in Table 4.4. Each block of the table represents different sample sizes. Since all the criteria agreed on the number of components (4,2,3 and 4 for n=100,500,2500,5000 respectively), the table only contains one set of parameter estimates for each sample size with the chosen number of components. It's surprising that when sample size is 100, the criteria all choose a 4 component HNM+NP model. Number 2,3,4 components actually has very similar means but a very different a value. The reason that a 4 component model was chosen could be the small size of the sample which leads to a unreliable result. It makes more sense that a 2 component HNM+NP model was chosen when n=500, and a 3 component model was chosen when n=2500, while a 4 component model was chosen when n=5000. Most of the birth weights are positively related to obstetric gestation, except the 4th component of the model when n=5000. The scatter plots and contour plots of the fitted densities are shown in Figure 4.1.

Notice that ρ is the correlation between X and Y calculated in Section 4.2.

$$\rho_{X,Y|M} = \frac{b_m \sigma_m}{\sqrt{\tau_m^2 + b_m^2 \sigma_m^2}}.$$

Table 4.4: Parameter Estimates for Models Chosen by AIC, BIC and sBIC

		Components m			
n	Parameter	m =1	m =2	m =3	m =4
n=100	μ	28.99998	39.04201	38.7438	37.90619
	a	-2142.667	-3839.172	-2375.976	-618.3162
	b	111.8333	176.212	152.952	121.1196
	τ^2	40379.93	40379.93	40379.93	40379.93
	σ^2	1.568323	1.568323	1.568323	1.568323
	π	0.0199999	0.4304674	0.4214359	0.1280968
	ρ	0.5717858	0.739383	0.6899711	0.6024646
n=500	μ	29.99576	38.89025		
	a	-3665.623	-2936.28		
	b	176.5592	161.011		
	τ^2	205661.8	205661.8		
	σ^2	1.896348	1.896348		
	π	0.03488106	0.9651189		
	ρ	0.4725081	0.4392326		
n=2500	μ	24.88911	34.5392	38.95195	
	a	-1751.01	-3534.945	-2483.193	
	b	103.15	168.4195	149.8614	
	τ^2	180511.9	180511.9	180511.9	
	σ^2	1.409501	1.409501	1.409501	
	π	0.005932702	0.05078797	0.9432793	
	ρ	0.276961	0.425822	0.3862633	
n=5000	μ	27.07182	34.50198	38.886	39.39033
	a	-2047.432	-2607.702	-2850.295	9801.834
	b	112.6555	144.2196	158.9236	-145.9206
	τ^2	166660.4	166660.4	166660.4	166660.4
	σ^2	1.432124	1.432124	1.432124	1.432124
	π	0.008422917	0.05575579	0.8911222	0.04469912
	ρ	0.3135809	0.3893958	0.4222906	-0.3932817

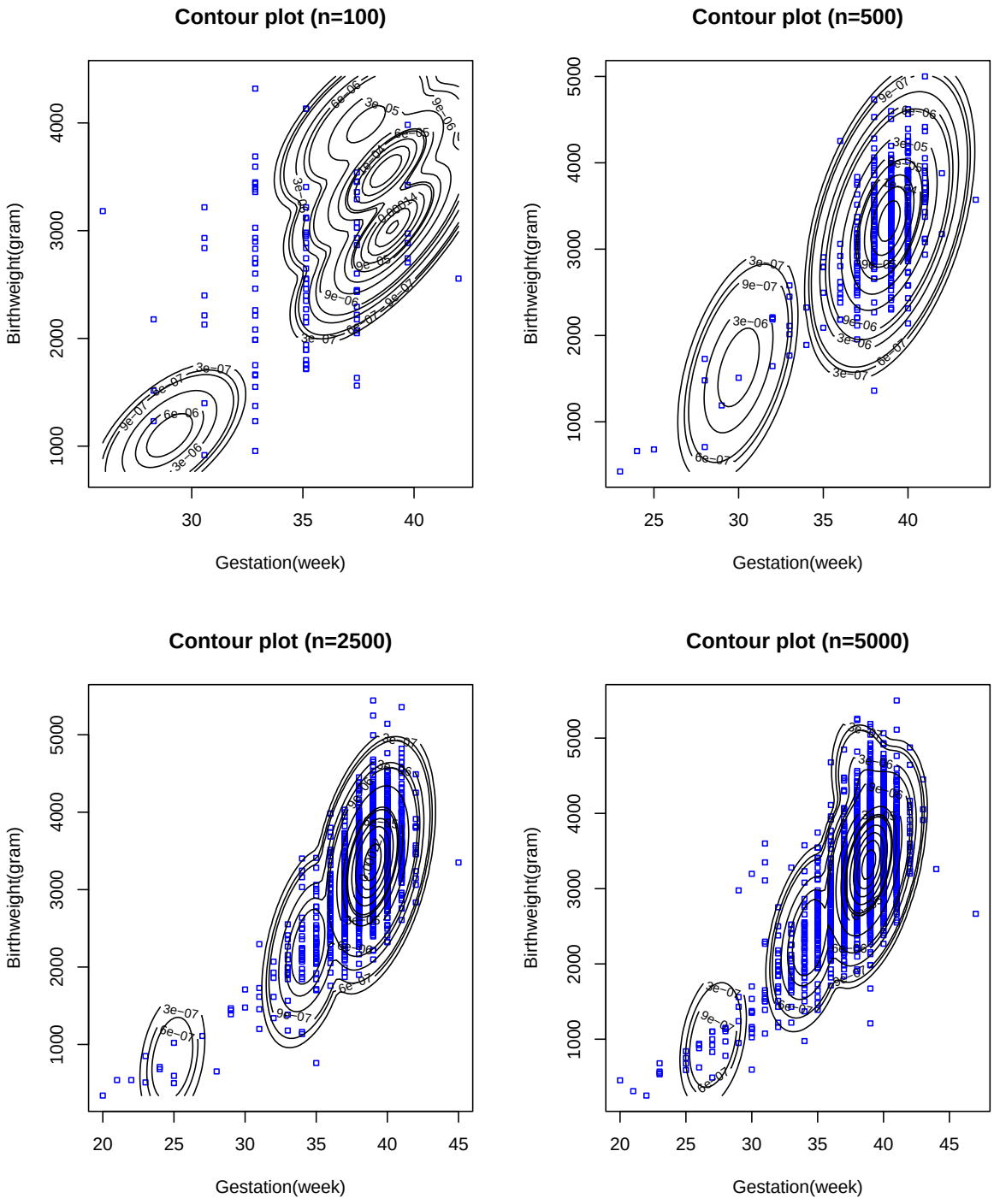

Figure 4.1: The Contour Plots of Fitted Density for Birth Weight Data

Chapter 5 A New Singular Information Criterion for HNM+NP Model

5.1 Motivation

In the literature of mixture modeling, one challenging and crucial topic is to determine the mixture complexity. A lot of methods have been used or developed in the past, like the AIC and BIC. In Section 4.3, we already discussed the problems AIC and BIC have when applied to choosing the number of mixture components. We also implemented a new information criterion for singular models by Drton (2013). In this Chapter we will develop a new data dependent information criterion inspired by Pilla and Charnigo's FLIC (2006). It is expected to work as least as good as sBIC, hopefully perform even better on estimating the complexity of HNM+NP model.

In this chapter, we will first define the new criterion. Then we will show the development of the criterion followed by the proof of its asymptotic properties. Some simulations will be done to see how well the criterion works. We will also compare this new criterion with AIC, BIC and sBIC by applying it to the birth weight data.

5.2 Multivariate Analysis of Variance - MANOVA

Before we construct the new criterion, we want to first review the multivariate analysis of variance - MANOVA. The new criterion will depend on a term that is inspired by MANOVA. MANOVA is a generalization of ANOVA when there is more than one dependent variable (Tabachnick and Fidell, 2001). Let Y_{iuj} represent the observation from ith outcome measurement of uth subject in jth group, $i = 1, ..., p$, $u = 1, ..., n_j$ and $j = 1, ..., J$. We want to decompose Total SSCP Matrix into Between SSCP and Within SSCP. SSCP stands for sum of squares and cross products. Use $\boldsymbol{T}, \boldsymbol{W}, \boldsymbol{B}$ to

represent each and suppose for illustration p=2, then

$$\boldsymbol{T} := \boldsymbol{W} + \boldsymbol{B}.$$

The sum of squares within groups can be calculated as follows:

$$\boldsymbol{W} = \sum_{j=1}^{J} \boldsymbol{W_j},$$

$$\boldsymbol{W_j} = \begin{bmatrix} SS_{j11} & SS_{j12} \\ SS_{j21} & SS_{j22} \end{bmatrix},$$

where

$$SS_{j11} = \sum_{u=1}^{n_j} (Y_{1uj} - \bar{Y}_{1\cdot j})^2$$

$$SS_{j22} = \sum_{u=1}^{n_j} (Y_{2uj} - \bar{Y}_{2\cdot j})^2$$

$$SS_{j12} = SS_{j21} = \sum_{u=1}^{n_j} (Y_{1uj} - \bar{Y}_{1\cdot j})(Y_{2uj} - \bar{Y}_{2\cdot j}).$$

Notice here if we divide $\boldsymbol{W_j}$ by $n_j - 1$, one can get

$$\frac{\boldsymbol{W_j}}{n_j - 1} = \begin{bmatrix} \hat{Var}_j(Y_1) & \hat{Cov}_j(Y_1, Y_2) \\ \hat{Cov}_j(Y_1, Y_2) & \hat{Var}_j(Y_2) \end{bmatrix}.$$

Thus

$$\boldsymbol{W} = \sum_{j=1}^{J} \boldsymbol{W_j} = \sum_{j=1}^{J} \begin{bmatrix} SS_{j11} & SS_{j12} \\ SS_{j21} & SS_{j22} \end{bmatrix}.$$

The sum of squares between groups can be calculated as follows:

$$\boldsymbol{B} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$$

where b_{ij} represents the sum of squares or cross products in the univariate case,

$$b_{11} = \sum_{j=1}^{J} n_j (\bar{Y}_{1\cdot j} - \bar{\bar{Y}}_{1\cdot\cdot})^2$$

$$b_{22} = \sum_{j=1}^{J} n_j (\bar{Y}_{2\cdot j} - \bar{\bar{Y}}_{2\cdot\cdot})^2$$

$$b_{12} = b_{21} = \sum_{j=1}^{J} n_j (\bar{Y}_{1\cdot j} - \bar{\bar{Y}}_{1\cdot\cdot})(\bar{Y}_{2\cdot j} - \bar{\bar{Y}}_{2\cdot\cdot}).$$

So the total SSCP matrix can be written as

$$\boldsymbol{T} = \boldsymbol{W} + \boldsymbol{B} = \sum_{j=1}^{J} \begin{bmatrix} SS_{j11} & SS_{j12} \\ SS_{j21} & SS_{j22} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{J} SS_{j11} + b_{11} & \sum_{j=1}^{J} SS_{j12} + b_{12} \\ \sum_{j=1}^{J} SS_{j21} + b_{21} & \sum_{j=1}^{J} SS_{j22} + b_{22} \end{bmatrix}.$$

5.3 Singular Flexible Information Criterion

We now know how to derive the between and within group variability for MANOVA. We draw an analogy in the mixture content. More specifically we consider $\boldsymbol{W}$ to be characterized by mixture component membership. Replace Y_1 with X, the obstetric gestation, and Y_2 with Y, the birth weight of new born infants, then the quantities listed below are analogous to those from MANOVA, with N representing the total sample size:

$$SS_{j11} = N\hat{\pi}_j \hat{\sigma}_j^2$$

$$SS_{j22} = N\hat{\pi}_j \hat{\tau}_j^2$$

$$SS_{j12} = SS_{j21} = N\hat{\pi}_j \hat{\sigma}_j \hat{\tau}_j$$

$$b_{11} = N \sum_{j=1}^{J} \hat{\pi}_j (\hat{\mu}_j - \bar{X})^2$$

$$b_{22} = N \sum_{j=1}^{J} \hat{\pi}_j (\hat{a}_j + \hat{b}_j \hat{\mu}_j - \bar{Y})^2$$

$$b_{12} = b_{21} = N \sum_{j=1}^{J} \hat{\pi}_j (\hat{\mu}_j - \bar{X})(\hat{a}_j + \hat{b}_j \hat{\mu}_j - \bar{Y})$$

Note that $N\hat{\pi}_j = \hat{n}_j$, the estimated number of persons in component j.

Then the within and between-component SSCP matrix becomes,

$$\boldsymbol{W} = \begin{bmatrix} \sum_{j=1}^{J} SS_{j11} & \sum_{j=1}^{J} SS_{j12} \\ \sum_{j=1}^{J} SS_{j21} & \sum_{j=1}^{J} SS_{j22} \end{bmatrix} = N \begin{bmatrix} \sum_{j=1}^{J} \hat{\pi}_j \hat{\sigma}_j^2 & \sum_{j=1}^{J} \hat{\pi}_j \hat{\sigma}_j \hat{\tau}_j \\ \sum_{j=1}^{J} \hat{\pi}_j \hat{\sigma}_j \hat{\tau}_j & \sum_{j=1}^{J} \hat{\pi}_j \hat{\tau}_j^2 \end{bmatrix}$$

$$\boldsymbol{B} = N \begin{bmatrix} \sum_{j=1}^{J} \hat{\pi}_j (\hat{\mu}_j - \bar{X})^2 & \sum_{j=1}^{J} \hat{\pi}_j (\hat{\mu}_j - \bar{X})(\hat{a}_j + \hat{b}_j \hat{\mu}_j - \bar{Y}) \\ \sum_{j=1}^{J} \hat{\pi}_j (\hat{\mu}_j - \bar{X})(\hat{a}_j + \hat{b}_j \hat{\mu}_j - \bar{Y}) & \sum_{j=1}^{J} \hat{\pi}_j (\hat{a}_j + \hat{b}_j \hat{\mu}_j - \bar{Y})^2 \end{bmatrix}$$

$$\boldsymbol{T} = N \begin{bmatrix} \sum_{j=1}^{J} \hat{\pi}_j (\hat{\sigma}_j^2 + (\hat{\mu}_j - \bar{X})^2) & \sum_{j=1}^{J} \hat{\pi}_j (\hat{\sigma}_j \hat{\tau}_j + (\hat{\mu}_j - \bar{X})(\hat{a}_j + \hat{b}_j \hat{\mu}_j - \bar{Y})) \\ \sum_{j=1}^{J} \hat{\pi}_j (\hat{\sigma}_j \hat{\tau}_j + (\hat{\mu}_j - \bar{X})(\hat{a}_j + \hat{b}_j \hat{\mu}_j - \bar{Y})) & \sum_{j=1}^{J} \hat{\pi}_j (\hat{\tau}_j^2 + (\hat{a}_j + \hat{b}_j \hat{\mu}_j - \bar{Y})^2) \end{bmatrix}$$

The determinants can be calculated accordingly,

$$|\boldsymbol{W}| = N^2 (\sum_{j=1}^{J} \hat{\pi}_j \hat{\sigma}_j^2)(\sum_{j=1}^{J} \hat{\pi}_j \hat{\tau}_j^2) - N^2 (\sum_{j=1}^{J} \hat{\pi}_j \hat{\sigma}_j \hat{\tau}_j)^2$$

$$|\boldsymbol{B}| = N^2 (\sum_{j=1}^{J} \hat{\pi}_j (\hat{\mu}_j - \bar{X})^2)(\sum_{j=1}^{J} \hat{\pi}_j (\hat{a}_j + \hat{b}_j \hat{\mu}_j - \bar{Y})^2) - N^2 (\sum_{j=1}^{J} \hat{\pi}_j (\hat{\mu}_j - \bar{X})(\hat{a}_j + \hat{b}_j \hat{\mu}_j - \bar{Y}))^2$$

$$\begin{aligned} |\boldsymbol{T}| =& N^2 (\sum_{j=1}^{J} \hat{\pi}_j (\hat{\sigma}_j^2 + (\hat{\mu}_j - \bar{X})^2))(\sum_{j=1}^{J} \hat{\pi}_j (\hat{\tau}_j^2 + (\hat{a}_j + \hat{b}_j \hat{\mu}_j - \bar{Y})^2)) \\ & - N^2 (\sum_{j=1}^{J} \hat{\pi}_j (\hat{\sigma}_j \hat{\tau}_j + (\hat{\mu}_j - \bar{X})(\hat{a}_j + \hat{b}_j \hat{\mu}_j - \bar{Y})))^2 \end{aligned}$$

then define the fraction of within-component to total variability of the HNM+NP model, averaged over candidate models, to be

$$\Lambda(X, Y) := \frac{1}{M} \sum_{j=1}^{M} \frac{|\boldsymbol{W}_j|}{|\boldsymbol{T}_j|} \geq \frac{1}{M},$$

where $\boldsymbol{W}_j$ is the within variability assuming a j-component model and $\frac{|\boldsymbol{W}_j|}{|\boldsymbol{T}_j|}$ is always less than or equal to 1 and M is the largest model complexity under consideration.

To see this, suppose we have two matrices

$$\boldsymbol{A} = \begin{vmatrix} a_1 & a \\ a & a_2 \end{vmatrix}, \quad \boldsymbol{B} = \begin{vmatrix} b_1 & b \\ b & b_2 \end{vmatrix},$$

and $\det(\boldsymbol{A}) = a_1 a_2 - a^2$, $\det(\boldsymbol{B}) = b_1 b_2 - b^2$. By the property of determinants, the determinant of $\boldsymbol{A} + \boldsymbol{B}$ is

$$\begin{aligned}
\det(\boldsymbol{A}+\boldsymbol{B}) &= \begin{vmatrix} a_1+b_1 & a+b \\ a+b & a_2+b_2 \end{vmatrix} = \begin{vmatrix} a_1+b_1 & a+b \\ a & a_2 \end{vmatrix} + \begin{vmatrix} a_1+b_1 & a+b \\ b & b_2 \end{vmatrix} \\
&= \begin{vmatrix} a_1 & a \\ a & a_2 \end{vmatrix} + \begin{vmatrix} b_1 & b \\ a & a_2 \end{vmatrix} + \begin{vmatrix} a_1 & a \\ b & b_2 \end{vmatrix} + \begin{vmatrix} b_1 & b \\ b & b_2 \end{vmatrix} \\
&= \det(\boldsymbol{A}) + b_1 a_2 + a_1 b_2 - 2ab + \det(\boldsymbol{B}).
\end{aligned}$$

If view matrix $\boldsymbol{A}$ as $\boldsymbol{W}_j$ and $\boldsymbol{B}$ as $\boldsymbol{B}_j$, then

$$\begin{aligned}
b_1 a_2 + a_1 b_2 - 2ab &= \hat{\pi}_j^2 \hat{\sigma}_j^2 (\hat{a}_j + \hat{b}_j \hat{\mu}_j - \bar{Y})^2 + \hat{\pi}_j^2 \hat{\tau}_j^2 (\hat{\mu}_j - \bar{X})^2 \\
&\quad - 2\hat{\pi}_j^2 \hat{\sigma}_j \hat{\tau}_j (\hat{\mu}_j - \bar{X})(\hat{a}_j + \hat{b}_j \hat{\mu}_j - \bar{Y}) \\
&= \hat{\pi}_j^2 [\hat{\sigma}_j (\hat{a}_j + \hat{b}_j \hat{\mu}_j - \bar{Y}) - \hat{\tau}_j (\hat{\mu}_j - \bar{X})]^2 \\
&\geq 0.
\end{aligned}$$

So $|\boldsymbol{T}_j| = |\boldsymbol{W}_j + \boldsymbol{B}_j| \geq |\boldsymbol{W}_j| + |\boldsymbol{B}_j|$ and $\frac{|\boldsymbol{W}_j|}{|\boldsymbol{T}_j|} \leq 1$. Further, note $\Lambda(X, Y) \geq \frac{1}{M}$. The larger $\Lambda(X, Y)$ is, the less heterogeneous the components are.

In Pilla and Charnigo (2006), they defined a bivariate ratio function, which is

$$g(n, \gamma) := \frac{\Phi[(\log(n))^\gamma] - \Phi(1)}{1 - \Phi(1)},$$

this function is defined for $n > \exp(1)$ and γ between M^{-1} and 1. Here $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution. This $g(n, \gamma)$ has some useful properties that can be used to develop the new information criterion.

Firstly, $g(n, \gamma)$ will increase as n or γ increases. Secondly, $g(n, \gamma)$ goes to 1 as n goes to infinity. Here, we can use this function as well.

Now define the penalty term

$$f(n, X, Y) = \log(n)^{g[n,\Lambda(X,Y)]}.$$

The *singular flexible information criterion* (SFIC) can be defined as

$$\text{SFIC}_j(X, Y) := \hat{l}(M_j) - \lambda_j(\pi_0)\log(n)^{g[n,\Lambda(X,Y)]} + [m(\pi_0) - 1](\log\log(n))^{g[n,\Lambda(X,Y)]}.$$

In the HNM+NP model, the multiplicity is anticipated to be 1, so the latter term disappears and SFIC becomes

$$\text{SFIC}_j(X, Y) := \hat{l}(M_j) - \lambda_j(\pi_0)\log(n)^{g[n,\Lambda(X,Y)]},$$

where $\hat{l}(M_j)$ is the maximized log likelihood function and M_j is drawn from a set of probability distributions with order j. The larger SFIC is, the better the model achieves balance between parsimony and fitting the sample data.

Notice that SFIC has the following properties,

1. The ratio of the penalty term $f(n, X, Y)$ to $\log(n)$ converges to 1 almost surely as n goes to infinity.

2. For finite n, if $\boldsymbol{B}$ and $\boldsymbol{W}$ suggest little heterogeneity, $\Lambda(X, Y)$ is large and thus $f(n, X, Y)$ is approximately close to $\log(n)$. In such case, the SFIC becomes close to "$\hat{l}(M_j) - \lambda_j(\pi_0)\log(n)$", which is the sBIC for HNM+NP model introduced in Chapter 4.

3. For finite n, if $\boldsymbol{B}$ and $\boldsymbol{W}$ suggest considerable heterogeneity, $\Lambda(X, Y)$ is small and thus $f(n, X, Y)$ is approximately close to 1. In such case, the SFIC reduces close to "$\hat{l}(M_j) - \lambda_j(\pi_0)$", which the penalty is $\lambda_j(\pi_0)$.

So the SFIC has a milder penalty than the sBIC when there is little heterogeneity between components.

(2) and (3) are easy to see. To show (1), we have

$$1 \geq \frac{f(n,X,Y)}{\log(n)} = \log(n)^{g[n,\Lambda(X,Y)]-1}$$
$$= \exp[(g[n,\Lambda(X,Y)]-1)\log\log(n)].$$

For a fixed value of $\Lambda(X,Y)$, call it $\Lambda^* > 0$, the rate of function $g[n,\Lambda^*]-1$ going to zero depends on the rate of $\log(n)^{\Lambda^*}$ going to infinity. Notice that

$$0 \geq g[n,\Lambda^*]-1 = \frac{\Phi(\log(n)^{\Lambda^*})-\Phi(1)}{1-\Phi(1)} - 1 = \frac{\Phi(\log(n)^{\Lambda^*})-1}{1-\Phi(1)}$$
$$\geq -\frac{\exp^{-(\log(n)^{2\Lambda^*}/2)}}{1-\Phi(1)} \to 0 \text{ as } n \to \infty.$$

Thus, $g[n,\Lambda^*]-1$ goes to zero faster than $\log\log(n)$ going to infinity, thus

$$\exp[(g[n,\Lambda^*]-1)\log\log(n)] \to 1 \text{ as } n \to \infty.$$

Since $\Lambda(X,Y) \geq \Lambda^* := \frac{1}{M}$,

$$\frac{f(n,X,Y)}{\log n} \to 1 \text{ a.s. as } n \to \infty.$$

5.4 Consistency of SFIC

To prove consistency, we can adapt Theorem 4.1 from Drton (2013). We already proved in Chapter 4 that the HNM+NP model satisfies assumptions (1)-(3), which are also assumptions (A1)-(A3) from Drton's paper. Since we only consider a finite set of models I, the desired conclusion will follow from pairwise comparisons of competing models.

Theorem 5.4.1. *(Consistency). Let $M_{\hat{i}}$ be the best model selected by SFIC, in other words,*

$$\hat{i} = \arg\max_{i \in I} SFIC(M_i).$$

Under assumptions (1)-(3) in Chapter 4,

$$P(M_i \text{ is a true model}) \to 1 \text{ as } n \to \infty.$$

Since the SFIC and sBIC have the same form of Bayes complexity for model M_j, which is $(\lambda_j(\pi_0), m_j(\pi_0))$, SFIC also satisfies that (1) asymptotically, the SFIC of a true model is larger than a false model, and (2) the SFIC of a true model M_j is maximized only when the Bayes complexity of that model is minimized among all true models. We establish these facts in our next two propositions.

Proposition 5.4.2. *Under assumption (2), the probability that SFIC(M_t) > SFIC(M_f) goes to 1 as $n \to \infty$, if M_t is a true model and M_f is a false model.*

Proof. Since SFIC and sBIC differ in the penalty term, we can adapt the proof of Drton's *Proposition 4.1* and conclude that

$$L'(M_f) = o_p(L'(M_t)),$$

$$\frac{L'(M_f)}{L'(M_t)} \xrightarrow{P} 0,$$

and hence

$$\frac{\exp(\text{sBIC}_f)}{\exp(\text{sBIC}_t)} \xrightarrow{P} 0.$$

By property 1 of SFIC,

$$\frac{\text{Penalty SFIC}_j}{\text{Penalty sBIC}_j} \xrightarrow{P} 1,$$

and hence

$$\frac{\hat{l}(M_j) - \log \text{Penalty SFIC}_j}{\hat{l}(M_j) - \log \text{Penalty sBIC}_j} \xrightarrow{P} 1,$$

where $\hat{l}(M_j)$ is the log likelihood function of HNM+NP model M_j. That is to say,

$$\frac{\text{SFIC}_j}{\text{sBIC}_j} \xrightarrow{P} 1.$$

We now claim that,

$$\log(n)^{g[n,\Lambda(X,Y)]} - \log(n) \xrightarrow{P} 0 \text{ as } n \to \infty.$$

Then we will have

$$\frac{\exp(\text{SFIC}_j)}{\exp(\text{sBIC}_j)} = \exp\{\lambda_j[\log(n)^{g[n,\Lambda(X,Y)]} - \log(n)]\} \xrightarrow{P} 1,$$

and then

$$\frac{\exp(\text{SFIC}_f)}{\exp(\text{SFIC}_t)} = \frac{\exp(\text{sBIC}_f)}{\exp(\text{sBIC}_t)} \cdot \frac{\exp(\text{SFIC}_f)}{\exp(\text{sBIC}_f)} \cdot \frac{\exp(\text{sBIC}_t)}{\exp(\text{SFIC}_t)} \xrightarrow{P} 0.$$

Therefore,

$$P[\exp(\text{SFIC}_f) > \exp(\text{SFIC}_t)] \to 0,$$

$$P[\text{SFIC}_f > \text{SFIC}_t] \to 0.$$

To see that the claim is true, notice that

$$1 - \Phi(t) < \exp(-\frac{t^2}{2}),$$

$$\Phi(\log(n)^\gamma) > 1 - \exp(-\frac{(\log(n)^\gamma)^2}{2}),$$

and

$$g[n, 1/M] > \frac{1 - \exp(-\frac{(\log(n)^{1/M})^2}{2}) - \Phi(1)}{1 - \Phi(1)} = 1 - \frac{\exp(-\frac{\log(n)^{2/M}}{2})}{1 - \Phi(1)},$$

Let $c_n := \log(n)$, then as $n \to \infty$, $c_n \to \infty$. Therefore

$$\log(n)^{g[n,1/M]} - \log(n) = c_n^{1 - \frac{\exp(-c_n^{2/M}/2)}{1-\Phi(1)}} - c_n = c_n(c_n^{-C\exp(-c_n^{2/M}/2)} - 1),$$

for some positive constant C. Then notice

$$c_n^{-C\exp(-c_n^{2/M}/2)} = \exp(-C\exp(-c_n^{2/M}/2)\log(c_n)) = \exp\left(-\frac{\log(c_n)}{C\exp(c_n^{2/M}/2)}\right),$$

then using Taylor's expansion gives us

$$\exp\left(-\frac{\log(c_n)}{C\exp(c_n^{2/M}/2)}\right) = 1 + \left(-\frac{\log(c_n)}{C\exp(c_n^{2/M}/2)}\right) + O\left(\left(\frac{\log(c_n)}{C\exp(c_n^{2/M}/2)}\right)^2\right),$$

$$c_n\left[\exp\left(\frac{\log(c_n)}{C\exp(c_n^{2/M}/2)}\right) - 1\right] = c_n\left[-\frac{\log(c_n)}{C\exp(c_n^{2/M}/2)} + O\left(\left(\frac{\log(c_n)}{C\exp(c_n^{2/M}/2)}\right)^2\right)\right]$$

$$= -\frac{c_n \log(c_n)}{C \exp(c_n^{2/M}/2)} + c_n O\left(\left(\frac{\log(c_n)}{C \exp(c_n^{2/M}/2)}\right)^2\right) \to 0 \text{ as } n \to \infty.$$

Finally, since $\Lambda(X, Y) \geq 1/M$, the same condition holds in probability when $\Lambda(X, Y)$ replace $1/M$. This completes the proof that $P(\text{SFIC}(M_t) > \text{SFIC}(M_f))$ goes to 1 as $n \to \infty$. □

Proposition 5.4.3. *Suppose M_s is a true model, that is, the data-generating distribution π_0 is in M_s, but M_s does not minimize Bayes complexity. Then assuming (1)-(3) are satisfied, the probability that a true model M_t that minimizes the Bayes complexity satisfies SFIC(M_t) > SFIC(M_s) goes to 1 as $n \to \infty$.*

Proof. Again, we can adapt Drton's *Proposition 4.2* and the proof of the previous Proposition 5.4.2 to conclude there exists a true model M_t such that

$$P(\text{SFIC}(M_t) > \text{SFIC}(M_s)) \to 1 \text{ as } n \to \infty.$$

□

5.5 Simulation Study

To see how the SFIC performs, we can conduct a simulation study. Similar to Section 4.5, 10 samples of size $n = 2500$ from four HNM+NP models will be randomly generated, each with 2,3,4 or 5 components, respectively. AIC, BIC, sBIC and SFIC will be calculated to do model selection. The parameters used for generating the random samples are listed in Table 5.1. The `R` code to generate this simulation study can be found in appendix.

Table 5.1: Parameter Combinations of HNM+NP Models for Simulation Study

M	2	3	4	5	M	2	3	4	5
π_1	0.4	0.2	0.2	0.1	a_1	-2	-2	-2	-2
π_2	0.6	0.5	0.3	0.2	a_2	2	1	-1	-1
π_3	-	0.3	0.3	0.3	a_3	-	2	1	0
π_4	-	-	0.2	0.2	a_4	-	-	2	1
π_5	-	-	-	0.2	a_5	-	-	-	2
μ_1	-2	-2	-2	-2	b_1	-2	-2	-2	-2
μ_2	2	-1	-1	-1	b_2	2	1	-1	-1
μ_3	-	2	1	0	b_3	-	2	1	0
μ_4	-	-	3	1	b_4	-	-	2	1
μ_5	-	-	-	2	b_5	-	-	-	2
σ^2	1	1	1	1	τ^2	2	2	2	2

Table 5.2 shows the results from simulation. When the true distribution is a 2 component normal mixture distribution, all criteria correctly choose the right number of components. Compare to the results from Section 4.5, this time all information criteria usually choose the right model when the true distribution is a 3 component normal mixture. In Section 4.5, all information criteria fail to correctly choose the 3 component normal mixture distribution, while this time 70% of the samples are correctly identified. This may be because the sample size increases from 1000 to 2500. As the sample size increases, the performance of all information criteria (except AIC, possibly) may improve.

Table 5.2: Number of Components Chosen by AIC, BIC, sBIC and SFIC, out of 10 samples of size 2500, from HNM+NP models in Table 5.1

	M=2				M=3				M=4				M=5			
Criterion	2	3	4	5	2	3	4	5	2	3	4	5	2	3	4	5
AIC	10	0	0	0	3	7	0	0	10	0	0	0	10	0	0	0
BIC	10	0	0	0	3	7	0	0	10	0	0	0	10	0	0	0
sBIC	10	0	0	0	3	7	0	0	10	0	0	0	10	0	0	0
SFIC	10	0	0	0	3	7	0	0	10	0	0	0	10	0	0	0

However, when the true distribution is a 4 or 5 component normal mixture, all criteria fail to choose the right number of components. Instead, they choose the 2 component normal mixture model. This may be due to the small differences between the component means when there are more than 3 components. When the component means occupy the same range but there are more components, the differences may be too small to identify and there may not be enough "power" to detect additional components. However, as the sample size goes up, the performance of information criteria (except AIC, possibly) should improve.

From the simulation results, we can say that SFIC does as good as the other information criteria.

5.6 Application to Vital Statistics Natality Birth Data Set

In Section 4.6, the HNM+NP model has been applied to the NCHS's Vital Statistics Natality Birth Data from 2011 to see the performance of sBIC compared to AIC and BIC. In this section, sFIC will be added and compare to the other three criteria. Detailed description of the data set can be found in Section 4.6.

Again, multiple sample sizes will be used (i.e. n $\in \{100, 500, 2500, 5000\}$). Within each sample size, 10 samples will be randomly drawn and each sample will be fitted to 2,3,4 and 5 component HNM+NP model via EM algorithm and AIC, BIC, sBIC, SFIC will be calculated, respectively. The obstetric gestation will be treated as the lower level and birth weight will be treated as the higher level since gestation might more reasonably be used to predict birth weight than vice versa. The number of components chosen by AIC, BIC, sBIC and SFIC are listed in Table 5.3.

Table 5.3: Numbers of Samples that Choose 2,3,4,5 Component Models

	AIC				BIC				sBIC				SFIC			
Sample Size	2	3	4	5	2	3	4	5	2	3	4	5	2	3	4	5
n=100	3	4	2	1	4	5	1	0	3	5	2	0	3	3	2	2
n=500	5	4	1	0	5	4	1	0	5	4	1	0	4	5	1	0
n=2500	1	3	3	3	1	3	3	3	1	3	3	3	1	3	3	3
n=5000	0	5	3	2	0	5	3	2	0	5	3	2	0	5	3	2

Table 5.3 displays the the results of each criterion applied to 10 random samples of the indicated size. When sample size is as small as 100, AIC, BIC and sBIC favor a 3- component model while SFIC has equal preference for both 2- and 3- component models. In this case, SFIC tends to choose more components than the others, weakening its preference for a 3-component model toward 4- and 5- component models. When the sample size increases to 500, the performances of all four criteria are similar. AIC, BIC, sBIC slightly favor a 2- component model, while SFIC slightly favor a 3- component model. Again, SFIC seems to be more likely to select more complicated models.

As the sample size increases to 2500, the behavior of all criteria tend to be more similar. They all result in the same number of components selected for each sam-

ple. Although 3-, 4- and 5- component models are more likely to be selected, the 10 samples are selected randomly and are a relatively small sub-sample compared to the population; we could have more variation among selected models if we investigated more samples. As for n=5000, models with more components are more likely to be chosen. None of the samples had selected a 2-component model by any criterion this time. Half of samples resulted in a 3-component model and some of the samples resulted in a 4- or 5- component model.

For small samples, AIC and SFIC tend to select a more complicated model than BIC and sBIC. As sample size increases, the resulting model may be a more complicated one as well. In either case, the outcomes are consistent with the our expectation based on the nature of the criteria and related theory.

5.7 Discussion

There have been a great number of studies on utilizing normal mixture models to clustering. McLachlan and Basford (1988) introduced a method to apply mixture models as a tool to do classification. These methods assume the true distribution of the population is a mixture distribution and each cluster is one mixture component. There are also `R` packages that can perform such tasks, for example, the `MCLUST` package by Fraley (2006), the mixtools package by Benaglia, Tatiana, et al. (2009). Currently, the BCN+NP model has been demonstrated on DNA microarray data sets. The model classifies the genes into three categories: no-, under- and over- differential expression.

However, the BCN+NP model has a limitation, the data may come from a 4 or more component normal mixture distribution or not a finite normal mixture at all! If this model can be extended to accommodate more components, it should have more flex-

ibility on fitting the data. EM algorithm is suggested to estimate the parameters.

The AIC, BIC, sBIC and SFIC can be used for model selection. Each criterion has its own advantages. The AIC has a penalty term depending on the number of parameters in the model, which can avoid underfitting. However, it's also well-known that the AIC is not a consistent estimator of the number of parameters of the correct model. On the other hand, the BIC is asymptotically consistent as an estimator, but it tends to choose a simpler model compared to the AIC which possibly causes underfitting of the model. The sBIC is an improved version of the BIC which accounts for singularity of the underlying model. In the case where singularity may exist, the sBIC is preferable. The SFIC is not only proved to be asymptotically consistent and accommodates singularity in the underlying model, it also has a data-driven penalty which makes it flexible for considering the sample size and taking into account the structure of the data. Data appearing more heterogeneous may yield a lighter penalty.

The other limitation of this study is the nuisance variance is assumed to be homoscedastic. When the unknown within-component variances are allowed to be heteroscedastic, convergence of the EM algorithm may become a problem, due to the large number of parameters that need to be estimated and the possible non-uniqueness of local extreme of the likelihood (Karlis and Xekalaki, 2003). Besides, heteroscedastic within-component variances may necessitate different hypothesis testing procedures for homogeneity or for two components versus more than two components. There is also the possibility of singularities (Chen and Gupta, 2010) in the likelihood depending on the parameter space, which will cause failure of the algorithm. More sophisticated methods can be employed or developed to handle unequal variances when naive application of the EM algorithm fails.

It is also of interest to investigate other probabilistic mixture models with or without hierarchical structures. In the Vital Statistics Natality Birth Data, the birth weight and obstetric gestation are always positive, so a hierarchical gamma or log-normal mixture model could be more appropriate. Wang, Yau and Lee (2002) applied a two-component hierarchical Poisson mixture regression model (HPMM) to the statewide obstetrical delivery data. They analyzed the inpatient length of stay (LOS) using the HPMM approach. Heydari and Amador-Jimenez (2012) compared Poisson, hierarchical Poisson-Gamma and hierarchical Poisson-Lognormal mixture models that have been implemented to estimate expected accident frequencies and identify hazardous sites which is a major concern of departments of transportation. Charnigo, Zhou and Dai (2013) developed a contaminated Chi-square mixture model that has been applied to DNA microarray data analysis. This model accommodates comparisons among three or more groups that yield strictly positive test statistics as in ANOVA. They investigated the genes that were related to aging and cognition. They also incorporated the procedure into a gene filtration process. These may be useful alternatives of our present model.

Appendices

A1

R code for simulation study in Section 2.5 (to count the number of samples that produce one and only one viable solution, displayed in Tables 2.1 - 2.4):

```
ViableSolutionsMoments<-function(m1,m2,m3,m4)
{

# Define the adjusted moments:

r1 <- m1
r2 <- m2-1
r3 <- m3-3*m1
r4 <- m4-6*m2+3

# Calculate coefficients of quintic equation for  \mu:

a0 <- r3^2*r4 - r4^2*r2
a1 <- -r3^3 + r4^2*r1
a2 <- -2*r1*r3*r4+ 2*r4*r2^2
a3 <- 2*r1*r3^2 - 2*r4*r1*r2
a4 <- r1^2*r4-r2^3
a5 <- -r1^2*r3 + r2^2*r1

# Find roots of quintic equation for \mu:

muroot <- polyroot(c(a0,a1,a2,a3,a4,a5))

# Find implied solutions for  p, \gamma_1, \gamma_2:

# q0 <- r0
q1 <- r1/muroot
q2 <- r2/muroot^2
q3 <- r3/muroot^3
```

```
q4 <- r4/muroot^4

#proot  <- (q2-q4)/(q3-q1)

proot  <- sqrt((q4-q3)/(q2-q1))

gamma1root <- (q2+proot*q1)/(1+proot)
gamma2root <- (q2-q1)/(proot+proot^2)


# Ascertain number of viable solutions to method of moments equations:

Viable<-
(Im(gamma1root)<1e-04)*(Im(gamma2root)<1e-04)*(Re(gamma1root)>0)*
(Re(gamma2root)>0)*(Re(1-gamma1root-gamma2root)>0)*(Im(proot)<1e-04)*
(Re(proot)>0)*(Im(muroot)<1e-04)*(Re(muroot)>0)*( abs( r1 - muroot*
(gamma1root-proot*gamma2root) ) < 1e-04)*( abs( r2 - muroot^2*
(gamma1root+proot^2*gamma2root) ) < 1e-04)*( abs( r3 - muroot^3*
(gamma1root-proot^3*gamma2root) ) < 1e-04)*( abs( r4 - muroot^4*
(gamma1root+proot^4*gamma2root) ) < 1e-04)

return(list(Viable=Viable,gamma1root=gamma1root,gamma2root=
gamma2root,muroot=muroot,proot=proot,r1=r1,r2=r2,r3=r3,r4=r4))
}

SimulationStudy <- function( mu,p,gamma1,gamma2,samplesize,repetitions )
{

Viability<-matrix(NA,repetitions,5)
MuRoot<-matrix(NA,repetitions,5)
PRoot <-matrix(NA,repetitions,5)
Gamma1Root<-matrix(NA,repetitions,5)
Gamma2Root<-matrix(NA,repetitions,5)
R1<-rep(NA,repetitions)
```

```
R2<-rep(NA,repetitions)
R3<-rep(NA,repetitions)
R4<-rep(NA,repetitions)
ViolationNonreal<- matrix(NA,repetitions,5)
ViolationMoments<- matrix(NA,repetitions,5)
ViolationRanges<- matrix(NA,repetitions,5)

for (k in 1:repetitions)
{
comp <- runif(samplesize)
dep  <- rnorm(samplesize)
obs  <- (comp < gamma1)*(dep+mu)+(comp > 1-gamma2)*(dep-p*mu)+
(comp < 1-gamma2)*(comp > gamma1)*dep
m1<- mean(obs)
m2<- mean(obs^2)
m3<- mean(obs^3)
m4<- mean(obs^4)
Object <- ViableSolutionsMoments(m1,m2,m3,m4)
Viability[k,]<-Object$Viable
MuRoot[k,]<-Object$muroot
PRoot[k,]<-Object$proot
Gamma1Root[k,]<-Object$gamma1root
Gamma2Root[k,]<-Object$gamma2root
R1[k]<-Object$r1
R2[k]<-Object$r2
R3[k]<-Object$r3
R4[k]<-Object$r4
ViolationNonreal[k,]<- pmax(abs(Im(Gamma1Root[k,]))>1e-04,
abs(Im(Gamma2Root[k,]))>1e-04,abs(Im(PRoot[k,]))>1e-04,
abs(Im(MuRoot[k,]))>1e-04)
ViolationMoments[k,]<- pmax(abs(R1[k]-MuRoot[k,]*Gamma1Root[k,]+
PRoot[k,]*MuRoot[k,]*Gamma2Root[k,])>1e-04,
 abs(R2[k]-MuRoot[k,]^2*Gamma1Root[k,]-PRoot[k,]^2*MuRoot[k,]^2*
Gamma2Root[k,])>1e-04,
```

```
 abs(R3[k]-MuRoot[k,]^3*Gamma1Root[k,]+PRoot[k,]^3*MuRoot[k,]^3*
Gamma2Root[k,])>1e-04,
 abs(R4[k]-MuRoot[k,]^4*Gamma1Root[k,]-PRoot[k,]^4*MuRoot[k,]^4*
Gamma2Root[k,])>1e-04)
ViolationRanges[k,]<- pmax( Re(Gamma1Root[k,])<0, Re(Gamma2Root[k,])<0,
Re(Gamma1Root[k,]+Gamma2Root[k,])>1, Re(MuRoot[k,])<0, Re(PRoot[k,])<0 )
}

#############
ViaRowSum=apply(Viability,1,sum)
Viability=cbind(Viability,ViaRowSum)
ViaColSum=apply(Viability,2,sum)
NumRep=c(1000,1000,1000,1000,1000,5000)
ViaColAvg=ViaColSum/NumRep
Viability=rbind(Viability,ViaColSum,ViaColAvg)

NonrealRowSum=apply(ViolationNonreal,1,sum)
ViolationNonreal=cbind(ViolationNonreal,NonrealRowSum)
NonrealColSum=apply(ViolationNonreal,2,sum)
NumRep=c(1000,1000,1000,1000,1000,5000)
NonrealColAvg=NonrealColSum/NumRep
ViolationNonreal=rbind(ViolationNonreal,NonrealColSum,NonrealColAvg)

MomentsRowSum=apply(ViolationMoments,1,sum)
ViolationMoments=cbind(ViolationMoments,MomentsRowSum)
MomentsColSum=apply(ViolationMoments,2,sum)
NumRep=c(1000,1000,1000,1000,1000,5000)
MomentsColAvg=MomentsColSum/NumRep
ViolationMoments=rbind(ViolationMoments,MomentsColSum,MomentsColAvg)

RangesRowSum=apply(ViolationRanges,1,sum)
ViolationRanges=cbind(ViolationRanges,RangesRowSum)
RangesColSum=apply(ViolationRanges,2,sum)
NumRep=c(1000,1000,1000,1000,1000,5000)
```

```
RangesColAvg=RangesColSum/NumRep
ViolationRanges=rbind(ViolationRanges,RangesColSum,RangesColAvg)
############

return(list(Viability=Viability,MuRoot=MuRoot,PRoot=PRoot,Gamma1Root=
Gamma1Root,Gamma2Root=Gamma2Root,R1=R1,R2=R2,R3=R3,R4=R4,
ViolationNonreal=ViolationNonreal,ViolationMoments=ViolationMoments,
ViolationRanges=ViolationRanges)
)}
```

A2

```
R code for simulation study in Section 3.4 (to estimate type I
error probability and power under different parameter combinations,
displayed in Tables 3.1 - 3.6):

BCNwNP.simul.T=function(nsize, nrepeat, t.df, mu1, mu2, g1, g2,
alpha1, alpha2, alpha3, alpha4){

resultn=resulta=matrix(NA, 1, nrepeat)
m1n=m2n=m3n=m4n=m5n=m6n=matrix(NA, 1, nsize)
m1a=m2a=m3a=m4a=m5a=m6a=matrix(NA, 1, nsize)
Tn=Un=Vn=Wn=matrix(NA, 1, nsize)
Ta=Ua=Va=Wa=matrix(NA, 1, nsize)

for (i in 1:nrepeat){
comp=runif(nsize)
tstat=rt(nsize, t.df)
zstatn=qnorm(pt(tstat, t.df))
gen=(comp < g1)*(tstat+mu1)+(comp > 1-g2)*(tstat+mu2)+(comp < 1-g2)*
(comp > g1)*tstat
zstata=qnorm(pt(gen, t.df))
```

```
m1n[i] = mean(zstatn)
m2n[i] = mean(zstatn^2)
m3n[i] = mean(zstatn^3)
m4n[i] = mean(zstatn^4)
m5n[i] = mean(zstatn^5)
m6n[i] = mean(zstatn^6)

m1a[i] = mean(zstata)
m2a[i] = mean(zstata^2)
m3a[i] = mean(zstata^3)
m4a[i] = mean(zstata^4)
m5a[i] = mean(zstata^5)
m6a[i] = mean(zstata^6)

  Tn[i] = sqrt(nsize/m2n[i])*m1n[i]
  Un[i] = sqrt(3*nsize/8)*(m4n[i]/(3*m2n[i]^2)-1)
  Vn[i] = sqrt(5*nsize/136)*(m6n[i]/(15*m2n[i]^3)-1)
  Wn[i] = sqrt(nsize/(15*m2n[i]^3))*m3n[i]

  Ta[i] = sqrt(nsize/m2a[i])*m1a[i]
  Ua[i] = sqrt(3*nsize/8)*(m4a[i]/(3*m2a[i]^2)-1)
  Va[i] = sqrt(5*nsize/136)*(m6a[i]/(15*m2a[i]^3)-1)
  Wa[i] = sqrt(nsize/(15*m2a[i]^3))*m3a[i]

resultn[i]=(abs(Tn[i])<qnorm(1-alpha1/2))*(abs(Un[i])<qnorm(1-alpha2/2))*
(abs(Vn[i])<qnorm(1-alpha3/2))*(abs(Wn[i])<qnorm(1-alpha4/2))
resulta[i]=(abs(Ta[i])<qnorm(1-alpha1/2))*(abs(Ua[i])<qnorm(1-alpha2/2))*
(abs(Va[i])<qnorm(1-alpha3/2))*(abs(Wa[i])<qnorm(1-alpha4/2))
  #1 not reject, 0 reject
 }
type1err=1-sum(resultn)/nrepeat
power=1-sum(resulta)/nrepeat
return(list(type1err,power))
}
```

```
BCNwNP.simul.Z=function(nsize, nrepeat, t.df, mu1, mu2, g1, g2,
alpha1, alpha2, alpha3, alpha4){

resultn=resulta=matrix(NA, 1, nrepeat)
m1n=m2n=m3n=m4n=m5n=m6n=matrix(NA, 1, nsize)
m1a=m2a=m3a=m4a=m5a=m6a=matrix(NA, 1, nsize)
Tn=Un=Vn=Wn=matrix(NA, 1, nsize)
Ta=Ua=Va=Wa=matrix(NA, 1, nsize)

for (i in 1:nrepeat){
comp=runif(nsize)
zstatn=rnorm(nsize)
zstata=(comp < g1)*(zstatn+mu1)+(comp > 1-g2)*(zstatn+mu2)+
    (comp < 1-g2)*(comp > g1)*zstatn

m1n[i] = mean(zstatn)
m2n[i] = mean(zstatn^2)
m3n[i] = mean(zstatn^3)
m4n[i] = mean(zstatn^4)
m5n[i] = mean(zstatn^5)
m6n[i] = mean(zstatn^6)

m1a[i] = mean(zstata)
m2a[i] = mean(zstata^2)
m3a[i] = mean(zstata^3)
m4a[i] = mean(zstata^4)
m5a[i] = mean(zstata^5)
m6a[i] = mean(zstata^6)

  Tn[i] = sqrt(nsize/m2n[i])*m1n[i]
```

```
  Un[i] = sqrt(3*nsize/8)*(m4n[i]/(3*m2n[i]^2)-1)
  Vn[i] = sqrt(5*nsize/136)*(m6n[i]/(15*m2n[i]^3)-1)
  Wn[i] = sqrt(nsize/(15*m2n[i]^3))*m3n[i]

  Ta[i] = sqrt(nsize/m2a[i])*m1a[i]
  Ua[i] = sqrt(3*nsize/8)*(m4a[i]/(3*m2a[i]^2)-1)
  Va[i] = sqrt(5*nsize/136)*(m6a[i]/(15*m2a[i]^3)-1)
  Wa[i] = sqrt(nsize/(15*m2a[i]^3))*m3a[i]

resultn[i]=(abs(Tn[i])<qnorm(1-alpha1/2))*(abs(Un[i])<qnorm(1-alpha2/2))*
(abs(Vn[i])<qnorm(1-alpha3/2))*(abs(Wn[i])<qnorm(1-alpha4/2))
resulta[i]=(abs(Ta[i])<qnorm(1-alpha1/2))*(abs(Ua[i])<qnorm(1-alpha2/2))*
(abs(Va[i])<qnorm(1-alpha3/2))*(abs(Wa[i])<qnorm(1-alpha4/2))
  #1 not reject, 0 reject
 }
type1err=1-sum(resultn)/nrepeat
power=1-sum(resulta)/nrepeat
return(list(type1err,power))
}
```

A3

R code for parameter estimation of Down's syndrome data set in Section 3.5 (Table 3.7):

```
install.packages("lattice",lib="C://Users//FAN//Dropbox//R//Rlib")
library(lattice,lib.loc="C://Users//FAN//Dropbox//R//Rlib")
install.packages("nlme",lib="C://Users//FAN//Dropbox//R//Rlib")
library(nlme,lib.loc="C://Users//FAN//Dropbox//R//Rlib")

############ apply to down syndrome data
############### function for calculating Z statistics
mixedDown <- function(y){
x<-       c(1,1,1,1,0,0,0,0,0,0,0)
```

```
Subject<- c(1,2,3,4,5,5,6,6,7,7,8)
Object1 <- lme( y ~ x, random = ~1 | Subject)
Tstat <- Object1$coef$fixed[2]/sqrt(Object1$varFix[2,2])
Zstat <- qnorm(pt(Tstat,6))
Zstat}

OptimChrom=function(data, startval, lower, upper){
############ calculate m1-m5
data=as.matrix(data[,c(9,10,11,12,20,21,22,23,24,25,26)])
temp=rep(NA,nrow(data))
for (i in 1:nrow(data)){
temp[i]=try(mixedDown(data[i,]))}
Down.Zstat=na.omit(as.numeric(temp))

m1=mean(Down.Zstat)
m2=mean(Down.Zstat^2)
m3=mean(Down.Zstat^3)
m4=mean(Down.Zstat^4)
m5=mean(Down.Zstat^5)

############### equation system
fn2=function(p){
f=(p[1]*p[2]+p[3]*p[4]-m1)^2+
(p[5]+p[1]^2*p[2]+p[3]^2*p[4]-m2)^2+
((p[1]^3+3*p[5]*p[1])*p[2]+(p[3]^3+3*p[5]*p[3])*p[4]-m3)^2+
(3*p[5]^2+(p[1]^4+6*p[5]*p[1]^2)*p[2]+(p[3]^4+6*p[5]*p[3]^2)*p[4]-m4)^2+
((p[1]^5+10*p[1]^3*p[5]+15*p[1]*p[5]^2)*p[2]+(p[3]^5+10*p[3]^3*p[5]+
15*p[3]*p[5]^2)*p[4]-m5)^2+max(0,(p[2]+p[4]-0.9))
f
}

############## results
resultmat1=matrix(NA,nrow=nrow(startval),ncol=12)
for (i in 1:nrow(startval)){
```

```
temp=optim(startval[i,], lower=lower, upper=upper,
fn2, hessian=TRUE,method="L-BFGS-B")
resultmat1[i,1]=temp$value
resultmat1[i,2:6]=temp$par
resultmat1[i,7:11]=round(eigen(temp$hessian)$values,4)
resultmat1[i,12]=temp$convergence
}
result1=resultmat1[order(resultmat1[,1]),]

###outcome
result=result1[1:20,]
return(result1)
}

############randomly generate starting value
mu1=runif(1000,0,3)
mu2=runif(1000,-3,0)
g1=runif(1000,0,1)
g2=runif(1000,0,1)
var=runif(1000,.1,1)
startv=cbind(mu1,g1,mu2,g2,var)
startval=startv[which(startv[,2]+startv[,4]<1),]
lower=c(0,0,-3,0,.1)
upper=c(3,1,0,1,1)

chrom0<-which(DownData[,29]==0);  chromosome0<-DownData[chrom0,];
Down1<-OptimChrom(chromosome0,startval,lower,upper)
......
chromy<-which(DownData[,29]=="y");  chromosomey<-DownData[chromy,];
Down25<-OptimChrom(chromosomey,startval,lower,upper)

############# p-value for Omnibus test ############
Omnibus.pvalue=function(data){
data=as.matrix(data[,c(9,10,11,12,20,21,22,23,24,25,26)])
```

```
temp=rep(NA,nrow(data))
for (i in 1:nrow(data)){
temp[i]=try(mixedDown(data[i,]))}
Down.Zstat=na.omit(as.numeric(temp))
n=length(Down.Zstat)

rho1=0.25
rho2=0.25
rho3=0.25
rho4=1-rho1-rho2-rho3

m1=mean(Down.Zstat)
m2=mean(Down.Zstat^2)
m3=mean(Down.Zstat^3)
m4=mean(Down.Zstat^4)
m5=mean(Down.Zstat^5)
m6=mean(Down.Zstat^6)
var=var(Down.Zstat)

T = sqrt(n/var)*m1
U = sqrt(3*n/8)*(m4/(3*m2^2)-1)
V = sqrt(5*n/136)*(m6/(15*m2^3)-1)
W = sqrt(n/(15*var^3))*m3

p1=2*(1-pnorm(abs(T)))
p2=2*(1-pnorm(abs(U)))
p3=2*(1-pnorm(abs(V)))
p4=2*(1-pnorm(abs(W)))

pvalue=min(1,p1/rho1,p2/rho2,p3/rho3,p4/rho4)

return(pvalue)
}
```

```
chrom0<-which(DownData[,29]==0);chromosome0<-DownData[chrom0,];
Down1<-Omnibus.pvalue(chromosome0)
......
chromy<-which(DownData[,29]=="y");chromosomey<-DownData[chromy,];
Down25<-Omnibus.pvalue(chromosomey)


############## Find MLE for BCN+NP model ############
shortCNNPfitMLE<-function(DATA,init1,init2,init3,useinit,n.repeat)
{
data=as.matrix(DATA[,c(9,10,11,12,20,21,22,23,24,25,26)])
temp=rep(NA,nrow(data))
for (i in 1:nrow(data)){
temp[i]=try(mixedDown(data[i,]))}
Down.Zstat=na.omit(as.numeric(temp))

DATA=Down.Zstat
n.size<-length(DATA)
n.repeat<-1
mu0<-0
sigma0<-var(DATA)
M<- 100
MM<- 100
tol<- 1e-5
nominal<- 0.05
C<- 0
K<- 10000
data<-matrix(DATA,nrow=n.size)
converge<-rep("Fail",n.repeat)
A<-D0<-lambda<-fail0<-fail3<-rep(0,n.repeat)
estimate<-matrix(ncol=n.repeat,nrow=3)
hat.sigma<-hat.alpha<-hat.mu<-matrix(ncol=n.repeat,nrow=K)
d<-hat.mu0<-hat.sigma0<-rep(0,n.repeat)
d.p<-d.mu<-d.sigma<-matrix(ncol=n.repeat,nrow=3)
```

```
###power.0<-power.3<-99
j<-1
hat.mu[1,j]<- 2*mean(DATA)
hat.alpha[1,j]<- 0.50
hat.sigma[1,j]<- 1
hat.mu0[j]<-0
hat.sigma0[j]<-mean((data[,j]-hat.mu0[j])^2)
if (useinit == 1)
{
hat.mu[1,j]<- init1
hat.alpha[1,j]<- init2
hat.sigma[1,j]<- init3
}
for (k in 2:K)
{
A<-hat.alpha[k-1,j]*(2*pi*hat.sigma[k-1,j])^(-0.5)*
exp(-(data[,j]-hat.mu[k-1,j])^2/(2*hat.sigma[k-1,j]))/
((1-hat.alpha[k-1,j])*(2*pi*hat.sigma[k-1,j])^(-0.5)*
exp(-(data[,j]-mu0)^2/(2*hat.sigma[k-1,j]))
+hat.alpha[k-1,j]*(2*pi*hat.sigma[k-1,j])^(-0.5)*
exp(-(data[,j]-hat.mu[k-1,j])^2/(2*hat.sigma[k-1,j])))
hat.alpha[k,j]<-( sum(A)+C ) / ( length(A) + 2*C )
hat.mu[k,j]<-sum(A*data[,j])/sum(A)
hat.sigma[k,j]<-mean((data[,j]-hat.mu[k,j])^2*A+(data[,j]-mu0)^2*(1-A))
if (hat.alpha[k,j]=="NaN" | hat.mu[k,j]=="NaN" | hat.sigma[k,j]=="NaN")
###if (is.nan(hat.alpha[k,j])==TRUE | is.nan(hat.mu[k,j])==TRUE
 | is.nan(hat.sigma[k,j])==TRUE)
{
converge[j]<-"NaN"
break
}
if (
sum(log(
((1-hat.alpha[k,j])*(2*pi*hat.sigma[k,j])^(-0.5)*
```

```
exp(-(data[,j]-mu0)^2/(2*hat.sigma[k,j]))+
hat.alpha[k,j]*(2*pi*hat.sigma[k,j])^(-0.5)*
exp(-(data[,j]-hat.mu[k,j])^2/(2*hat.sigma[k,j])))))-
sum(log(((1-hat.alpha[k-1,j])*(2*pi*hat.sigma[k-1,j])^(-0.5)*
exp(-(data[,j]-mu0)^2/(2*hat.sigma[k-1,j]))+
hat.alpha[k-1,j]*(2*pi*hat.sigma[k-1,j])^(-0.5)*
exp(-(data[,j]-hat.mu[k-1,j])^2/(2*hat.sigma[k-1,j])))))+
C*log(4*(1-hat.alpha[k,j])* hat.alpha[k,j])-
C*log(4*(1-hat.alpha[k-1,j])* hat.alpha[k-1,j] ) < tol)
{
converge[j]<-"Conv."
estimate[,j]<-c(hat.alpha[k,j],hat.mu[k,j],hat.sigma[k,j])
break
}
}
if (converge[j]=="NaN"  )
{estimate[,j]<-estimate[,j-1]}
if (converge[j]=="Fail")
{estimate[,j]<-c(hat.alpha[K,j],hat.mu[K,j],hat.sigma[K,j])}
d.p[1,j]<-(-1)
d.p[2,j]<-(1-estimate[1,j])
d.p[3,j]<-estimate[1,j]
d.mu[1,j]<-mu0
d.mu[2,j]<-mu0
d.mu[3,j]<-estimate[2,j]
d.sigma[1,j]<-hat.sigma0[j]
d.sigma[2,j]<-estimate[3,j]
d.sigma[3,j]<-estimate[3,j]
for (l in 1:3)
{
for (m in 1:3)
{
d[j]<-(d[j]+d.p[l,j]*d.p[m,j]*(2*pi*(d.sigma[l,j]+d.sigma[m,j]))^(-0.5)*
exp(-(d.mu[l,j]-d.mu[m,j])^2/(2*(d.sigma[l,j]+d.sigma[m,j]))))
```

```
}
}
D0[j]<-(4*sqrt(pi*hat.sigma0[j]))* d[j]*n.size
lambda[j]<-2*sum(log( (d.p[2,j]*(2*pi*d.sigma[2,j])^(-0.5)*
exp(-(data[,j]-d.mu[2,j])^2/(2*d.sigma[2,j]))+ d.p[3,j]*(2*pi*d.sigma[3,j])^(-0.5)
exp(-(data[,j]-d.mu[3,j])^2/(2*d.sigma[3,j])) )/
((2*pi*d.sigma[1,j])^(-0.5)*exp(-(data[,j]-d.mu[1,j])^2/(2*d.sigma[1,j])))))+
2*C*log(4*d.p[2,j]*d.p[3,j])
D0<-D0[j]
lambda<-lambda[j]
hatalpha<-d.p[3,j]
hatmu<-d.mu[3,j]
hatsigma<-d.sigma[3,j]
hatsigma0<-hat.sigma0[j]
modlik<-sum(log(
((1-hatalpha)*(2*pi*hatsigma)^(-0.5)*exp(-(data[,j]-mu0)^2/(2*hatsigma))+
hatalpha*(2*pi*hatsigma)^(-0.5)*exp(-(data[,j]-hatmu)^2/(2*hatsigma)))))+
C*log(4* 1-hatalpha * hatalpha)

return(list(hatalpha=hatalpha,hatmu=hatmu,hatsigma=hatsigma))
}

chrom0<-which(DownData[,29]==0);chromosome0<-DownData[chrom0,];
Down1<-shortCNNPfitMLE(chromosome0,0,0.5,1,0)
......
chromy<-which(DownData[,29]=="y");chromosomey<-DownData[chromy,];
Down25<-shortCNNPfitMLE(chromosomey,0,0.5,1,0)

############ Simulation study to find delta #############
Delta.Explore=function(data,alpha,mu,variance,delta){
nsize=nrow(data)
nrepeat=1000
result=rep(NA,nrepeat)
for (i in 1:nrepeat){
```

```
  comp=runif(nsize)
zstat=qnorm(comp,0,sqrt(var))
gen=(comp < alpha)*zstat+(comp > alpha)*(zstat+mu)

m1=mean(gen)
m2=mean(gen^2)
m3=mean(gen^3)
m4=mean(gen^4)
m5=mean(gen^5)
m6=mean(gen^6)
 sgm=rbind(c(m2-m1^2,m3-m1*m2,m4-m1*m3),
c(m3-m1*m2,m4-m2^2,m5-m2*m3),
c(m4-m1*m3,m5-m2*m3,m6-m3^2))
dh=c(6*m1*variance, 2*m2-2*variance, -m1)
crival=2*m2*delta+3*m1^2*delta+delta^2+qnorm(0.95)*sqrt(dh%*%sgm%*%dh/nsize)
    hstat=(m2-variance)^2+3*m1^2*variance-m1*m3
result[i]=hstat>crival # 1=reject null
}
return(sum(result))
}

############## p-value for Unilateral test #############
Pvalue.Unilateral=function(DATA,delta){
data=as.matrix(DATA[,c(9,10,11,12,20,21,22,23,24,25,26)])
temp=rep(NA,nrow(data))
for (i in 1:nrow(data)){
temp[i]=try(mixedDown(data[i,]))}
Zstat=na.omit(as.numeric(temp))
n=length(Zstat)

m1=mean(Zstat)
m2=mean(Zstat^2)
m3=mean(Zstat^3)
m4=mean(Zstat^4)
```

```
m5=mean(Zstat^5)
m6=mean(Zstat^6)
simvar=var(Zstat)
 sgm=rbind(c(m2-m1^2,m3-m1*m2,m4-m1*m3),
c(m3-m1*m2,m4-m2^2,m5-m2*m3),
c(m4-m1*m3,m5-m2*m3,m6-m3^2))
dh=c(6*m1*simvar, 2*m2-2*simvar, -m1)
 penalty=2*m2*delta+3*m1^2*delta+delta^2
 hstat=(m2-simvar)^2+3*m1^2*simvar-m1*m3
temp= (hstat-penalty)/sqrt(dh%*%sgm%*%dh/n)
pvalue=1-pnorm(abs(temp))
return(pvalue)
}

p0=Pvalue.Unilateral(chromosome0,0.15)
......
py=Pvalue.Unilateral(chromosomey,0.15)
```

A4

```
R code of simulation studies for hierarchical model.
########## random sample from HNMNP ##########
rHNMNP=function(n,mu,a,b,tau2,sigma2,p){
ii=findInterval(runif(n),cumsum(p))+1
x=rnorm(n,mu[ii],sqrt(sigma2[ii]))
y=rnorm(n,a[ii]+b[ii]*x,sqrt(tau2[ii]))
return(list(x=x,y=y,index=ii))
}

########## get the log likelihood density of HNMNP model ##########
HNMNP.llk=function(mu,a,b,tau2,sigma2,p, x, y){
n=length(x)
m=length(mu)
```

```
#calculate complete data loglik function
Wt=lt=llt=matrix(NA,n,m)
for (i in 1:n){
for (j in 1:m){
pt=p[j]
at=a[j]
bt=b[j]
mt=mu[j]
tt=tau2[j]
st=sigma2[j]
xt=x[i]
yt=y[i]
Wt[i,j]=pt*exp(-0.5*((xt-mt)^2/st+(yt-at-bt*xt)^2/tt))/(2*pi*sqrt(st*tt))
}}

llt=log(Wt)

if (m==1) {
llk=sum(llt)
W=rep(1,length(llk))} else {W=apply(Wt,1, function(x) x/sum(x))
llk=sum(t(W)*llt)}

return(list(W=W,loglik=llk))
}



########## get the estimates using EM algorithm ##########
HNMNP.EM=function(mu0,a0,b0,tau20,sigma20,p0,x,y){
n=length(x)
m=length(mu0)
differ=1

while(differ>0.0001){
tempw=HNMNP.llk(mu0,a0,b0,tau20,sigma20,p0,x,y)$W
```

```
w=t(tempw)

#update variables
newmu=newa=newb=newtau2=newsig2=newp=rep(NA,m)

for (j in 1:m){
wt=w[,j]
newp[j]=sum(wt)/n
newmu[j]=sum(wt*x)/sum(wt)
tempmat=matrix(c(sum(wt), sum(wt*x), sum(wt*x), sum(wt*x^2)),2,2)
temp=solve(tempmat)%*%rbind(sum(wt*y), sum(wt*x*y))
newa[j]=temp[1,]
newb[j]=temp[2,]
newsig2[j]=sum(wt*(x-newmu[j])^2)/sum(wt)
newtau2[j]=sum(wt*(y-newa[j]-x*newb[j])^2)/sum(wt)
}
newsig2=rep(sum(newp*newsig2),m)
newtau2=rep(sum(newp*newtau2),m)

#update results
newllk=HNMNP.llk(newmu,newa,newb,newtau2,newsig2,newp, x, y)$loglik
oldllk=HNMNP.llk(mu0,a0,b0,tau20,sigma20,p0, x, y)$loglik
differ=abs(newllk-oldllk)
a0=newa
b0=newb
tau20=newtau2
mu0=newmu
sigma20=newsig2
p0=newp
}#end of while loop

llk=HNMNP.llk(newmu,newa,newb,newtau2,newsig2,newp,x,y)$loglik

return(list(a=newa,b=newb,tau2=newtau2,mu=newmu,sigma2=newsig2,p=newp,loglik=llk))
```

```
}

########## get the learning coefficient for one HNMNP model ##########
sBIC.learncoef=function(maxm, nullm){
upperbound=0.5*(3*nullm+maxm+1)
return(upperbound)
}

########## get the log likelihood function vector ##########
HNMNP.llk.vec=function(maxm, x, y){
llkvec=lambdavec=muvec=avec=bvec=tau2vec=sig2vec=pvec=rep(NA, maxm)

ab=as.vector(glm(y~x)$coef)
mu5=fivenum(x)
varx=var(x)
vary=var(y)
n=length(x)

for (i in 1:maxm){
mu=seq(mu5[2], mu5[4], length.out=i)
a=rep(ab[1],i)
b=rep(ab[2],i)
tau2=rep(vary,i)
sigma2= rep(varx,i)
p=rep(1/i,i)
temp=HNMNP.EM(mu, a , b , tau2 , sigma2, p , x , y)
llkvec[i]=temp$loglik
wdet=n^2*sum(temp$p*temp$sigma2)*sum(temp$p*temp$tau2)-
n^2*(sum(temp$p*sqrt(temp$tau2*temp$sigma2)))^2
tdet=n^2*sum(temp$p*(temp$sigma2+(temp$mu-mean(x))^2))*
sum(temp$p*(temp$tau2+(temp$a+temp$b*temp$mu-mean(y))^2)
)-n^2*sum(temp$p*(sqrt(temp$tau2*temp$sigma2)+
(temp$mu-mean(x))*(temp$a+temp$b*temp$mu-mean(y)))
)^2
```

```
lambdavec[i]=wdet/tdet
}
lambda=mean(lambdavec)
return(list(llkvec=llkvec,lambda=lambda))
}

########## get the information criteria ##########
IC.func=function(maxm, minm, mult, x, y, loglikvec,lambda){
learnmat=rep(NA,maxm)
n=length(y)

for (j in minm:maxm){
learnmat[j]=sBIC.learncoef(maxm, j)
}

sBIC=loglikvec-learnmat*log(n)
AIC=-2* loglikvec[minm:maxm]+2*(4*seq(minm,maxm,1) +2-1)
BIC=-2* loglikvec[minm:maxm]+(4*seq(minm,maxm,1) +2-1)*log(n)
SFIC=loglikvec-learnmat*log(n)^((pnorm(log(n)^(lambda))-
pnorm(1))/(1-pnorm(1)))

return(list(AIC=AIC, BIC=BIC, sBIC=sBIC,SFIC=SFIC))
}
```

Bibliography

[1] Allison, B. D., Gadbury, G. L., Heo, M. *A mixture model approach for the analysis of microarray gene expression data*. ELSEVIER, Computational Statistics and Data Analysis 39 (2002) 1-20.

[2] Azaïs, Jean-Marc, Éisabeth Gassiat, and Cécile Mercadier. *Asymptotic distribution and local power of the log-likelihood ratio test for mixtures: bounded and unbounded cases*. Bernoulli 12.5 (2006): 775-799.

[3] Azaïs, Jean-Marc, Éisabeth Gassiat, and Cécile Mercadier. *The likelihood ratio test for general mixture models with possibly unknown structural parameter*. ESAIM: Probability and Statistics 13 (2009): 301-327.

[4] Benaglia, T., Chauveau, D., Hunter, D. R., Young, D. S. *mixtools: An R package for analyzing finite mixture models*. Journal of Statistical Software 2009, 32(6), 1-29.

[5] Bilmes, J. A. *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixed and Hidden Markov Models*. U.C. Berkely, TR-97-02, April 1998.

[6] Blehaut, H., Mircher, C., Ravel, A., Conte, M., De Portzamparc, V., Poret, G., ... Sturtz, F. G. *Effect of leucovorin (folinic acid) on the developmental quotient of children with Down's syndrome (trisomy 21) and influence of thyroid status*. PloS one, 5(1), e8394, 2010.

[7] Burnham, K. P., and David R. A.. *Multimodel inference understanding AIC and BIC in model selection*. Sociological methods research 33.2 (2004): 261-304.

[8] Casella, G. and Berger, R.L. *Statistical Inference*. Second Edition. Duxbury Press, 2002.

[9] Charnigo, R., Chesnut, L.W., LoBianco, T., Kirby R.S *Thinking outside the curve, Part I: Modeling birthweight distribution*. Biomedcentral Pregnancy and Childbirth, 10, Article 37. 2010.

[10] Charnigo, R., Fan, Q., Bittel, D., Dai, H. *Testing unilateral versus bilateral normal contamination*. Statistics Probability Letters, 83(1), 163-167, 2013

[11] Charnigo, R., Pilla, R. S. *Semiparametric mixtures of generalized exponential families*. Scandinavian Journal of Statistics, 34, 535-551. 2006.

[12] Charnigo, R., Sun, J. *Asymptotic relationships between the D-test and likelihood ratio-type tests for homogeneity*. Statistica Sinica, 20(2), 497, 2010.

[13] Charnigo, R., Sun, J. *Testing homogeneity in a mixture distribution via the L2 distance between competing models*. Journal of the American Statistical Association, 99 (466), 488-498, 2004.

[14] Charnigo, R., Zhou, F., Dai, H. *Contaminated Chi-Square Modeling and Large-Scale ANOVA Testing*. J Biomet Biostat, 4(157), 2, 2013

[15] Chen, H., Chen, J. *The Likelihood Ratio Test for Homogeneity in Finite Mixture Models*. The Canadian Journal of Statistics / La Revue Canadienne de Statistique, Vol. 29, No. 2 (Jun., 2001), pp. 201-215

[16] Chen, H., Chen, J., Kalbfleisch, J. D. *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, 2001.

[17] Chen, H., Chen, J., Kalbfleisch, J. D. *Testing for a finite mixture model with two components*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 66 (1), 95-115. 2004.

[18] Chen, J., Li, P. *Hypothesis test for normal mixture models: The EM approach*. The Annals of Statistics, 2523-2542, 2009.

[19] Chen, Y., Gupta, M. R. *EM demystified: An expectation-maximization tutorial*. University of Washington 2010.

[20] Conway, J. B., John B. *Functions of one complex variable*. Vol. 2. New York: Springer, 1973.

[21] Dacunha-Castelle D, Gassiat E. *Testing the order of a model using locally conic parameterization: Population mixtures and stationary ARMA processes*. The Annals of Statistics 1999, Vol 27, No. 4, 1178-1209.

[22] Dai, H., Charnigo, R. *Contaminated Normal Modeling with Application to Microarray Data Analysis*. The Canadian Journal of Statistics, Vol. 38, No. 3, 2010, pages 315-332.

[23] Dai, H., Charnigo, R. *Inferences in contaminated regression and density models*. Sankhya: The Indian Journal of Statistics, 842-869, 2008b.

[24] Dai, H., Charnigo, R. *Omnibus testing and gene filtration in microarray data analysis*. Journal of Applied Statistics, 35(1), 31-47, 2008a.

[25] Dempster, A. P., Laird, N. M. and Rubin, D. B. *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society: Series B (1977).

[26] Drton, M., and Plummer, M. *A Bayesian information criterion for singular models*. arXiv preprint arXiv:1309.0911 (2013).

[27] Fan, Q., Charnigo, R., Talebizadeh, Z., Dai, H. *Hypothesis Testing in Normal Admixture Models to Detect Heterogeneous Genetic Signals*. J Biomet Biostat, Volume 5, Issue 5, 2014.

[28] Fraley, C., Raftery, A. E. *MCLUST version 3: an R package for normal mixture modeling and model-based clustering*. WASHINGTON UNIV SEATTLE DEPT OF STATISTICS, 2006.

[29] Ghosh, J.K. and Sen, P.K. *On the asymptotic performance of the log likelihood ratio statistics for the mixture model and related results*. In Proceedings of the Berkeley Conference in Honor of J. Neyman and J. Kiefer (Lucien M. Le Cam and Richard A. Olshen, eds.) 2, 789-806. Wadsworth, 1985.

[30] Hartigan, J. A. *A failure of likelihood asymptotics for normal mixtures*. In proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer (Vol. 2, pp. 807-810). Wadsworth, Belmont, CA.

[31] Heydari, M., Amador-Jimenez, L. E. *Comparing Full Bayes Likelihoods to Predict Road Accidents and Identify Potential Hazardous Sites*. Journal of Civil Engineering and Science 2012, 1(3).

[32] Karlis, D., Xekalaki, E. *Choosing initial values for the EM algorithm for finite mixtures*. Computational Statistics Data Analysis 2003, 41(3), 577-590.

[33] Kendziorski, C.M., Newton, M.A., Lan, H., and Gould, M.N. *On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles*. Statistics in Medicine, 22, 3899-3914, 2003.

[34] Keribin, C. *Consistent estimation of the order of mixture models*. Sankhya: The Indian Journal of Statistics, Series A: 49-66, 2000.

[35] Lemdani, M., Pons, O. *Likelihood ratio tests in contamination models*. Bernoulli, 5(4), 705-719, 1999.

[36] Li, P., Chen, J. *Testing the order of a finite mixture*. Journal of the American Statistical Association, 105(491), 2010.

[37] Mao, R., Wang, X., Spitznagel, E., Frelin, L., Ting, J., Ding, H., Kim, J., Ruczinski, I., Downey, T., and Pevsner, J. *Primary and secondary transcriptional effects in the developing human Down syndrome brain and heart*. Genome Biology, 6, R107, 2005.

[38] McLachlan, G., Peel, D. *Finite Mixture Models*. Wiley, New York. 2000.

[39] McLachlan G.J., Basford K.E. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker 1988, New York.

[40] Parker, R. A. and Rothenberg, R. B. *Identifying Important Results from Multiple Statistical Tests*. Statistics in Medicine 17 (1988) 1031-1043.

[41] Pilla, R. S., Charnigo, R. *Consistent estimation and model selection in semiparametric mixtures*. 2006.

[42] Raudenbush, S. W., and Bryk, S. A., *Hierarchical linear models*. Thousand Oaks: Sage Publications, 2002.

[43] Redner, R. and Walker, H. *Mixture densities, maximum likelihood and the EM algorithm*. SIAM Review, Vol. 26, pp. 195-239, Apr. 1984.

[44] Roizen, N. J., Patterson, D. *Down's syndrome*. The Lancet, Volume 361, Issue 9365, 12 April 2003, Pages 1281-1289, ISSN 0140-6736.

[45] Schwarz, G. E. *Estimating the dimension of a model*. Annals of Statistics 6 (2): 461-464.

[46] Shaked, M. and Shanthikumar, J. G. *Stochastic Orders and their Applications*. Associated Press, 1994.

[47] Snijders, T. A. B., and Bosker, R. J., *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks: Sage Publications, 1999.

[48] Tabachnick, Barbara G., and Linda S. Fidell. *Using multivariate statistics*. Fifth Edition. Pearson, 2001.

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

[50] Wang, K., Yau, K. K., Lee, A.H. *A hierarchical Poisson mixture regression model to analyse maternity length of hospital stay*. Statistics in medicine 2002, 21(23), 3639-3654.

[51] Watanabe, S. *Algebraic geometry and statistical learning theory*. Volumn 25, Cambridge University Press, 2009.

[52] Wilcock, D. M. *Neuroinflammation in the aging down syndrome brain; lessons from Alzheimer's disease*. Current gerontology and geriatrics research, 2012.

Vita

Qian Fan was born in Suzhou, Jiangsu Province, China. After completing her school work at Suzhou No.3 High School in Suzhou in 2003, Qian entered Shanghai University in Shanghai, China. In 2009, Qian attended the University of Kentucky. She received a Bachelor of Business Administration from Shanghai University in May 2007, Master of Science with a major in statistics from the University of Kentucky in May 2011.

Qian Fan