Estimation in Mixture Models

by

Zhengmin Zhang, B.Eng., M.Sc.

A thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

School of Mathematics and Statistics
Ottawa-Carleton Institute for Mathematics and Statistics

Carleton University
Ottawa, Ontario, Canada
September, 2008

© Copyright
2008, Zhengmin Zhang
NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author’s permission.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

AVIS:
L’auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l’Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L’auteur conserve la propriété du droit d’auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n’y aura aucun contenu manquant.
Abstract

Mixture models arise in a wide variety of settings. This thesis is devoted to the estimation of mixing distributions, focusing on the maximum likelihood estimation and the method of moments estimation.

First of all, we discuss the identifiability in mixture models. We propose a sufficient condition for identifiability in the class of mixtures of multinomial distributions with a fixed number of components that implies Teicher's (1963) sufficient condition for identifiability in binomial mixtures with a fixed number of components. We also make a comparison between our result in terms of the number of components and Lindsay's (1995) result in terms of the index of mixing distributions.

In the maximum likelihood estimation, we describe the location of the maximum likelihood vector $\hat{L}$ from a geometric perspective. For self-consistency of the NPMLE, we provide conditions for the validity of the generalized likelihood equations in mixture distribution scenarios for which Lindsay's (1981) conditions are not accurate. We also discuss the uniqueness of the MLE; our result is quite striking since it asserts that if the discrete component mass function $f(x; \theta)$ is strictly totally positive and the Gateaux derivative $l'(\hat{L}, L(\theta)) \equiv 0$ in $\theta$, then for all $\theta$, $f(x; \theta)$ is a distribution on $\{y_1, \cdots, y_D\}$, the whole observation set. From this result, we can draw the conclusion that if the discrete component mass $f(x; \theta)$ is strictly totally positive and $f(x; \theta)$ is not supported on $\{y_1, \cdots, y_D\}$ for certain $\theta$, then the MLE of the mixing distribution is unique.

In the method of moments estimation, we reveal the relationship among the determination of the moment problem, the continuity (or discontinuity) of the underlying distribution function and the limit of the window function. By the method of moments, we construct a consistent estimate for the mixing distribution in a negative binomial
mixture, a normal mixture and a binomial mixture. We also prove the validity of Lindsay's (1989) modified moment method for mixing distributions supported on [0, 1] with the number of support points fixed. We demonstrate that the estimator from this method is weakly consistent and the convergence rate is $N^{-\frac{1}{2}}(\log \log N)^{\frac{1}{2}}$. 
Acknowledgements

I am very grateful to my supervisors, Professors A.K.Md.E. Saleh and P.J. Farrell for their guidance, encouragement, help and financial support.

Thanks are due to my committee: Professors H. Chen, C.G. Park, R. Balan and R. Thomas.

My thanks are also to Professors C.G. Park and M. Ould Haye for their teaching me courses and their concerns for my research.

I would like to thank Professor M. Csörgő for his help and generous financial support during the first two years of my Ph.D. study.

Finally I dedicate this thesis to my daughter Qixia with love.
Contents

Abstract ii

Acknowledgements iv

Contents v

1 Introduction and Summary 1

1.1 Introduction ........................................ 1

1.2 Summary of the Main Results .......................... 3

1.2.1 Nonparametric Maximum Likelihood Estimation .... 3

1.2.2 Method of Moments Estimation of Mixing Distributions .... 6

2 Identifiability of Mixing Distributions 9

2.1 Introduction ........................................ 9

2.2 Finite Mixtures of Multinomial Distributions .......... 12

2.2.1 Related Convex Geometry ......................... 12

2.2.2 Identifiability .................................. 15

2.3 Results on General Mixtures ........................ 28

3 Nonparametric Maximum Likelihood Estimates (NPMLEs) of Mixing Distributions 31
3.1 Introduction ................................................................. 31
3.2 Lindsay's NPMLE Theory and Its Supplement .......................... 33
  3.2.1 Introduction .......................................................... 33
  3.2.2 Fundamental Mixture NPMLE Theorems .......................... 35
  3.2.3 Duality Method of Solving for $\hat{\mathbf{L}}$ .................. 40
3.3 Convex Geometry ......................................................... 43
3.4 Location of the Maximum Likelihood Vector $\hat{\mathbf{L}}$ .............. 51
3.5 Self-Consistency of NPMLEs in Mixture Models ...................... 61
3.6 Other Developments on MLEs for Mixing Distributions .............. 70
  3.6.1 Consistency ......................................................... 70
  3.6.2 Binomial Mixtures ................................................. 71

4 Uniqueness of the NPMLE for a Mixing Distribution .................. 73
  4.1 Introduction .......................................................... 73
  4.2 Totally Positive Kernels ............................................. 75
  4.3 Preliminary Lemmas .................................................. 77
  4.4 Uniqueness of the NPMLE for a Mixing Distribution .............. 83

5 Method of Moments Estimation of Mixing Distributions ............... 88
  5.1 Introduction .......................................................... 88
  5.2 The Problem of Moments ............................................. 91
    5.2.1 Determination of the Problem of Moments, Continuity (or Dis-
         continuity) of the Underlying Distribution Function and the Limit
         of Window Function .............................................. 91
    5.2.2 Preliminary Lemmas ............................................. 96
  5.3 Method of Moments Estimation of Mixing Distributions .......... 100
Chapter 1

Introduction and Summary

1.1 Introduction

Mixture models are encountered in many contexts and for some situations may be more realistic than regular parametric models for describing many real life problems. For example, a mixture of Poisson distributions may be used to model the distribution of insurance claims during a fixed period, while a mixture of a family of exponential distributions can be considered as a possible model for a life time distribution. We elaborate on these two examples for illustration respectively.

Regarding the first example, suppose that insurance claims from a given insurant over a fixed time may be assumed to follow a Poisson distribution, but the parameter will vary from insurant to insurant. In such a case, the parameter $\lambda$ can be viewed as a random variable with distribution $Q$ and the probability of exactly $n$ insurance claims during the fixed period of time is given by

$$p(n) = \int_0^\infty e^{-\lambda} \frac{\lambda^n}{n!} dQ(\lambda).$$

So far as the second example is concerned, assume that the life span of any given individual in a species has an exponential distribution where the parameter $\beta$ varies
with living environmental conditions. If \( \beta \) has distribution \( Q \), then the probability that a random individual in the species has a life span less than or equal to \( x \) is given by

\[
\int_0^x \int_0^\infty \frac{1}{\beta} e^{-t/\beta} \, dQ(\beta) \, dt,
\]

where the probability density associated with \( x \) is

\[
\int_0^\infty \frac{1}{\beta} e^{-x/\beta} \, dQ(\beta).
\]

Abstracting from these two examples we have the following general definition for a mixture distribution.

Let \( \{f(x; \phi), \phi \in \Omega\} \) be a parametric family of density (or mass) functions with respect to some \( \sigma \)-finite measure \( \mu \) and \( Q(\phi) \) be a probability distribution function on \( \Omega \). Then

\[
f(x; Q) = \int_{\Omega} f(x; \phi) \, dQ(\phi)
\]

(1.1)
is called the mixture density, \( f(x; \phi) \) the component density and \( Q \) the mixing distribution.

When the distribution \( Q \) degenerates to a one point distribution (with mass one to take some value \( \phi \)), the mixture density (1.1) degenerates to a parametric density. So a mixture model is more general than a parametric model.

Assume the underlying random variable of \( f(x; Q) \) be \( X \). In mixture (1.1), if we treat \( \phi \) as a random variable with sample space \( \Omega \) and distribution function \( Q \), then \( f(x; \phi) \) is the conditional density of \( X \) given \( \phi \), and the mixture density \( f(x; Q) \) is the marginal density of \( X \) in the random pair \((X, \phi)\), which has the joint density \( f(x; \phi) \, dQ(\phi) \). Since only random variable \( X \) is observable, all statistical inference about \( Q \) is based only on the mixture density \( f(x; Q) \) and the sample \( X_1, \cdots, X_N \) from it.
Clearly, in order that such statistical inference can be meaningfully considered, the mixture (1.1) should be identifiable, that is, the mapping

\[ M : Q \rightarrow f(x; Q) \]

should be one to one. In the literature, identifiability has been considered by numerous researchers including Teicher (1961, 1963), Yokowitz and Spragins (1968), Blum and Susarla (1977), Al-Hussaini and Ahmad (1981), Ahmad (1988) and Lindsay (1995). Our discussion on identifiability in Chapter 2 will mainly focus on finite mixtures of multinomial distributions. We partly extend a result of Teicher (1963) and compare it with Lindsay's (1995) result.

1.2 Summary of the Main Results

This thesis is devoted to the study of estimation in mixture models. It consists of two relatively independent parts, but with the theme of studying estimates of mixing distributions. The first part is devoted to nonparametric maximum likelihood estimates (NPMLE) of mixing distributions, and is the focus of Chapters 3 and 4. The second part is concerned with the method of moments estimation in mixture models, and is presented in Chapters 5 and 6. In the next two subsections, we summarize the results of these four chapters.

1.2.1 Nonparametric Maximum Likelihood Estimation

Chapter 3 begins with a review of Lindsay's (1995) NPMLE theory, and gives detailed proofs for some of his results.

The location of the maximum likelihood vector is crucial in solving for the NPMLE of a mixing distribution. Our main contribution as to NPMLE is Theorem 3.8A, which
gives the exact location of the maximum likelihood vector on the boundary of the convex hull of the unicomponent likelihood curve. Specifically, this theorem states the following.

The maximum likelihood vector \( \hat{L} \) is a unique element in \( \mathcal{M} \) which belongs to the intersection of all open half spaces

\[
\{ p : H_{p_0}(p) > 0 \}, \quad p_0 \neq \hat{L}, \quad p_0 \in \mathcal{M},
\]

that is,

\[
\cap_{p_0 \in \mathcal{M}} \{ p : H_{p_0}(p) > 0 \} \cap \mathcal{M} = \{ \hat{L} \}.
\]

Herein, \( \mathcal{M} \) is the convex hull of the unicomponent likelihood curve, \( H_{p_0}(p) = 0 \) is the hyperplane passing through \( p_0 \) and orthogonal to the gradient vector \( \nabla l(p_0) \) of the likelihood function \( l \) at \( p_0 \), and \( H_{p_0}(p) > 0 \) is the open half space with boundary \( H_{p_0}(p) = 0 \) and containing \( p_0 + \nabla l(p_0) \). Similar to Theorem 3.8A, Theorems 3.8B and 3.8C also describe the location of the maximum likelihood vector \( \hat{L} \).

In Theorem 3.11, we propose a sufficient condition for the validity of the generalized likelihood equation under mixture \( f(x; Q) \), which is different from that in Lindsay’s (1981) original theorem. We give an example to illustrate that Lindsay’s (1981) condition cannot justify his proof. Theorem 3.11 is stated as follows.

Suppose that for any \( x \) in the sample space, \( \ln f(x; \phi) \) is differentiable with respect to \( \phi \) in \( \Omega \). Suppose further that \( g(\phi) \) is a differentiable function and for any \( t \) in a small neighborhood of 0, \( 1 + tg'(\phi) > \varepsilon \), for some small positive number \( \varepsilon \). If \( \hat{Q} \) is an NPMLE for the mixing distribution, then

\[
\sum_{i=1}^{n} E \left[ g(\phi) \cdot \frac{\partial \ln f(X; \phi)}{\partial \phi} \bigg| X = x_i; \hat{Q} \right] = 0.
\]
Moreover if \( f(x; \phi) \) is twice differentiable in \( \phi \) over \( \Omega \), then

\[
\sum_{i=1}^{n} E \left[ \frac{\partial^2 f(X; \phi)}{f(X; \phi)} \cdot g^2(\phi) - \frac{\partial \ln f(X; \phi)}{\partial \phi} \cdot g'(\phi)(1 - g(\phi)) \bigg| X = x_i; \hat{Q} \right] \\
\leq \sum_{i=1}^{n} \left\{ E \left[ g(\phi) \cdot \frac{\partial \ln f(X; \phi)}{\partial \phi} \bigg| X = x_i; \hat{Q} \right] \right\}^2.
\]

Chapter 4 discusses the uniqueness of the NPMLE of a mixing distribution. We give the limits of the range of the set of zeroes of the Gateaux derivative \( l'(\hat{L}, L(\theta)) \) in Lemma 4.5, which states the following.

If \( f(x; \theta) \) is STP, then

(a) For all \( \hat{\theta}_i \in Z, \hat{\theta}_1 \leq \hat{\theta}_i \leq \hat{\theta}_D \).

(b) If \( l'(\hat{L}, L(\hat{\theta}_1)) = 0 \), then \( \hat{\theta}_1 = u, y_1 = \inf \mathcal{X} \). If \( l'(\hat{L}, L(\hat{\theta}_D)) = 0 \), then \( \hat{\theta}_D = v, y_D = \sup \mathcal{X} \).

In addition, we assert in Theorem 4.2 that when the discrete mass function \( f(x; \theta) \) is strictly totally positive (STP),

\[
l'(\hat{L}, L(\theta)) \equiv 0
\]

implies that \( f(x; \theta) \) is supported on \( \{y_1, \cdots, y_D\} \) and

\[
\hat{L} = \hat{p}.
\]

Specifically, Theorem 4.2 states the following.

Let \( f(x; \theta) \) be STP, then

(a) \( l'(\hat{L}, L(\theta)) = 0 \)

for all \( \theta \) if and only if

\[
f(y; Q^0) = \hat{p}(y)
\]
for all \( Q^0 \in Q \), i.e., \( Q^0 \) such that

\[
L(Q^0) = \hat{L}.
\]

(b) If

\[
l'(\hat{L}, L(\theta)) = 0
\]

for all \( \theta \) and \( f(x; \theta) \) is a discrete distribution, then \( f(x; \theta) \) is supported on \( \{y_1, y_2, \ldots, y_D\} \) for all \( \theta \).

The uniqueness theorem in Simar (1976) is a by-product of our Theorem 4.2 (or Corollary 4.1).

1.2.2 Method of Moments Estimation of Mixing Distributions

Chapter 5 begins with the study of the \((n+1)\)-point distribution with one support point given that fits the first \( 2n+1 \) prescribed moments \( m_0, m_1, \ldots, m_{2n} \). By the Tchebycheff inequality given in Ahiezer (1965) and Lindsay and Basak (2000), in Theorem 5.1 and 5.2 we reveal the relationship among the determination of the moment problem, the continuity (or discontinuity) of the underlying distribution and the limit of the window function. Theorem 5.1 is summarized as follows.

If the solution \( \varphi \) of the moment problem

\[
m_n = \int_{-\infty}^{\infty} x^n d\varphi(x), \quad n = 0, 1, \ldots
\]

is continuous, then \( \varphi \) is completely determined by \( m \) if and only if

\[
w_n(x; m) \to 0
\]

for each \( x \in R \).

Herein,

\[
w_n(x; m) = \frac{1}{v'_n(x) M^{-1} v_n(x)},
\]
where $M_n$ is the moment matrix

$$M_n = \begin{pmatrix} m_0 & m_1 & \cdots & m_n \\ m_1 & m_2 & \cdots & m_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ m_n & m_{n+1} & \cdots & m_{2n} \end{pmatrix}$$

and $v_n(x) = (1, x, \cdots, x^n)'$. Theorem 5.2 states the following.

If $\varphi$ is completely determined by $m$, then

$$\varphi(x_0) - \varphi(x_0 - 0) = c(x_0),$$

where

$$c(x_0) = \lim_{n \to \infty} w_n(x_0; m)$$

For estimation in mixture models, we discuss the identifiability of a negative binomial mixture and a normal mixture. By the method of moments, we construct a consistent nonparametric estimate for the mixing distribution in a negative binomial mixture, a normal mixture and a binomial mixture. Using a result of Lindsay (1989), we simplify Tucker's proof of weak convergence.

In Chapter 6, we study the method of moments estimation in finite mixtures where the mixing distribution is supported on $[0,1]$ and the number of components is known. We prove the validity of the modification method proposed by Lindsay (1989) and we illustrate the virtue of his modification in comparison with the truncation method. Specifically, in Lindsay's method, the corrected estimator $\tilde{G}_N$ fits the first $2n+1$ or $2n$ moments, while the corrected estimator $G_N^*$ by truncation may not fit any of the moments. In addition, in solving for $\tilde{G}_N$ we only need to solve a polynomial equation of degree $n-1$ or $n$, while in solving for $G_N^*$, we have to solve a polynomial equation of degree $n+1$. Thus, the computational burden is reduced. In Chapter 6, we also demon-
strate that the modified estimator obtained by Lindsay's (1989) method is weakly consistent, and that Lindsay's modified estimator has convergence rate \( N^{-\frac{1}{2}} (\log \log N)^{\frac{1}{2}} \), which is not much slower than the rate \( N^{-\frac{1}{2}} \) of the MLE established by Chen (1995).
Chapter 2
Identifiability of Mixing Distributions

2.1 Introduction

Let $f(x; \phi), \phi \in \Omega$ be a parametric density (or mass) family. In order to estimate the distribution function $Q$ in the mixture density

$$f(x; Q) = \int f(x; \phi) dQ(\phi)$$

based on the sample from $f(x; Q)$, an important question that must be addressed is the following. When the parametric component density family $f(t; \phi), \phi \in \Omega$ and the mixture density $f(t; Q)$ are given, can we determine the mixing distribution $Q$ uniquely? This problem is related to the following concept.

**Definition 2.1.** The mixture

$$f(t; Q) = \int f(t; \phi) dQ(\phi)$$

is called identifiable if

$$\int f(t; \phi) dQ(\phi) = \int f(t; \phi) dQ'(\phi)$$
implies that the mixing distributions \( Q \) and \( Q' \) are same, i.e., \( Q = Q' \).

We denote

\[ \mathcal{H} = \{ f(t, Q) : Q \in \mathcal{Q} \} , \]

where \( \mathcal{Q} \) is a set of distribution functions on \( \Omega \). We say the class \( \mathcal{H} \) of mixtures is identifiable if each mixture in \( \mathcal{H} \) is identifiable.

Identifiable mixture models seem to be ideal regarding inference problems. A valuable study of identifiability of mixtures has been initiated by Teicher (1961, 1963). Other important literature on the identifiability of mixtures include Yakowitz and Spragins (1968), Blum and Susarla (1977), Al-Hussaini and Ahmad (1981), Ahmad (1988) and Lindsay (1995).

Teicher (1961) and Blum and Susarla (1977) gave some general results for the identifiability of mixing distributions by resorting to Fourier transforms and the Riesz representation theorem, respectively.

Teicher (1963) used a class of transforms to give a condition under which the class of all finite mixtures of the given distribution family is identifiable. Consequently, a large class of finite mixtures are proved to be identifiable. Applying Teicher's (1963) result, Al-Hussaini and Ahmad (1981) showed that the classes of finite mixtures of ten families of distributions are identifiable. Further, Ahmad (1988) proved the identifiability of finite mixtures of another five distribution families through utilizing the improved version of Teicher's (1963) result by Chandra (1977). We will discuss these works in detail in section 2.3, and provide the two lists of distributions there.

In this chapter, we mainly deal with multinomial mixtures. Let \( f(\phi), \phi \in \Omega \) be a
one parameter multinomial density family with sample space \( \{0, 1, \ldots, T\} \), that is,

\[
f(\phi) = \begin{pmatrix} f(0; \phi) \\ f(1; \phi) \\ \vdots \\ f(T; \phi) \end{pmatrix},
\]

where \( f(i; \phi) \) is the probability of the underlying random variable taking a value of \( i \) when the parameter is \( \phi \).

One of the most important such families is the binomial density family \( \text{Bin}(T, \theta), \theta \in [0, 1] \), with sample space \( \{0, 1, \ldots, T\} \). For fixed \( \theta \in [0, 1] \), the density vector is the \( T + 1 \)-dimensional vector

\[
\begin{pmatrix} \binom{T}{0}(1 - \theta)^T \\ \binom{T}{1}\theta(1 - \theta)^{T-1} \\ \vdots \\ \binom{T}{T}\theta^T \end{pmatrix}.
\]

We consider the multinomial mixture

\[
f(Q) = \int_\Omega f(\phi)dQ(\phi) = \sum_{i=1}^m \pi_i f(\phi_i),
\]

where \( m \) is fixed and \( \pi_i \geq 0, \sum_{i=1}^m \pi_i = 1 \).

For any \( \phi \in \Omega \), we call \( f(\phi) \) the component density vector. Clearly, the mixing distribution \( Q \) has the support points \( \phi_1, \ldots, \phi_m \) and the corresponding weights \( \pi_1, \ldots, \pi_m \). Moreover, if we define the matrix

\[
F = [f(\phi_1), f(\phi_2), \ldots, f(\phi_m)] = \begin{pmatrix} f(0; \phi_1) & f(0; \phi_2) & \cdots & f(0; \phi_m) \\ f(1; \phi_1) & f(1; \phi_2) & \cdots & f(1; \phi_m) \\ \vdots & \vdots & \ddots & \vdots \\ f(T; \phi_1) & f(T; \phi_2) & \cdots & f(T; \phi_m) \end{pmatrix}
\]

and the vector

\[
\pi = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_m \end{pmatrix},
\]
then equation (2.1) can be rewritten as a matrix equation:

$$f(Q) = F\pi.$$  \hspace{1cm} (2.2)

Our goal is to estimate the unknown mixing distribution $Q$ in the mixture model (2.1), or equivalently to estimate the support points $\phi_1, \ldots, \phi_m$, and the weights $\pi_1, \ldots, \pi_m$. We refer to $\phi_i$ and $\pi_i$, $i = 1, \ldots, m$ as component and mixing parameters respectively. Sometimes we denote $(\phi_1, \ldots, \phi_m)'$ by $\phi$.

We propose a sufficient condition in Theorem 2.1 for the class of mixtures (2.1) to be identifiable that includes the sufficient condition in Proposition 4 (i) in Teicher (1963) as a special case. In Example 2.1, we demonstrate this to be the case.

In addition, we prove the extension of Proposition 5 in Lindsay (1995, page 45) about a sufficient condition for the class of mixtures (2.1) to be identifiable, but where $\phi_i$ is known, $\pi_i > 0$ and $f(t; \phi), \ t = 0, \ldots, T$ constitute a Tchebycheff system. In Lindsay's (1995) proof, $f(t; \phi)$ is the binomial family, which is a special case of a Tchebycheff system family. We also prove Proposition 6 (our Theorem 2.3) in Lindsay (1995, page 48) about a necessary and sufficient condition for the class of mixtures (2.1) to be identifiable where $f(t; \phi)$ is a Tchebycheff system density. Again in Lindsay's proof, $f(t; \phi)$ is the binomial family. Finally, we make a comparison between Theorem 2.1 and Theorem 2.3.

\textbf{2.2 Finite Mixtures of Multinomial Distributions}

\textbf{2.2.1 Related Convex Geometry}

In this section, we introduce the relevant convex geometry that is helpful for understanding mixture models and analyzing the identifiability of multinomial mixtures.
Definition 2.2. A subset $M$ of $R^n$ is called an affine set if

$$(1 - \lambda)x + \lambda y \in M$$

for every $x \in M$, $y \in M$ and $\lambda \in R$. (A synonym for "affine set" used by other authors is "affine manifold").

Subspaces of $R^n$ are affine sets that contain the origin. Each non-empty affine set is parallel to a unique subspace. The dimension of a non-empty affine set is defined as the dimension of the subspace parallel to it.

Definition 2.3. A $(n - 1)$-dimensional affine set in $R^n$ is called a hyperplane.

The intersection of an arbitrary collection of affine sets is again affine. Therefore, given any $S \subseteq R^n$ there exists a unique smallest affine set containing $S$. This set is called the affine hull of $S$ and is denoted by $\text{aff}(S)$.

Definition 2.4. A set of $m + 1$ points $b_0, b_1, \cdots, b_m$ is said to be affinely independent if $\text{aff}(b_0, b_1, \cdots, b_m)$ is $m$-dimensional.

The points $b_0, b_1, \cdots, b_m$ are affinely independent if and only if

$$b_1 - b_0, \cdots, b_m - b_0$$

are linearly independent.

Definition 2.5. A subset $M$ of $R^n$ is said to be convex if

$$(1 - \lambda)x + \lambda y \in M$$

whenever $x \in M$, $y \in M$ and $0 \leq \lambda \leq 1$. 
A subset of $\mathbb{R}^n$ is convex if and only if it contains all the convex combinations of its elements. The intersection of an arbitrary collection of convex sets is convex.

**Definition 2.6.** The intersection of all convex sets containing a given subset $S$ of $\mathbb{R}^n$ is called the convex hull of $S$ and is denoted by $\text{conv}(S)$.

The convex hull $\text{conv}(S)$ of $S$ is the unique smallest convex set containing $S$.

For any $S \subset \mathbb{R}^n$, $\text{conv}(S)$ consists of all the convex combinations of the elements of $S$. Actually, it suffices to consider convex combinations involving $n+1$ or fewer elements at a time. This important refinement is known as the Caratheodory Theorem.

**Lemma 2.1.** (Caratheodory Theorem) Given a set $S$ in $\mathbb{R}^n$, every element $x$ in $\text{conv}(S)$ can be expressed as a convex combination of at most $n + 1$ elements of $S$. If $x$ is in the boundary of $\text{conv}(S)$, $n + 1$ can be replaced by $n$.

This famous theorem is very useful in the study of mixture models, especially in the study of the nonparametric maximum likelihood estimator of a mixing distribution.

**Definition 2.7.** A set which is the convex hull of finitely many points is called a polytope. If $\{b_0, b_1, \cdots, b_m\}$ is affinely independent, its convex hull is called an $m$-dimensional simplex, and $b_0, b_1, \cdots, b_m$ are called the vertices of the simplex.

The dimension of a convex set $M$ means the dimension of the affine hull of $M$.

Of particular interest to us is the case when the vectors are multinomial density vectors, that is, the case where the vectors have nonnegative entries which sum to one. Such vectors constitute the probability simplex

$$P^T = \{ p : p'1 = 1, \ p \geq 0 \},$$
a convex set with dimension $T$ in $\mathbb{R}^{T+1}$. In fact,

$$ P^T = \text{conv}\{e_0, \ldots, e_T\}, $$

where $e_j$ is the $(T+1)$-dimensional vector with $(j+1)$-th coordinate being one and all the others being zero.

We examine $P^2$, the 2-dimensional probability simplex in $\mathbb{R}^3$, as shown in figure 2.1. Inside $P^2$, there are three multinomial component density vectors $f_1$, $f_2$ and $f_3$. The set of mixture density vectors of them

$$ M := \left\{ f(Q) : f(Q) = \pi_1 f_1 + \pi_2 f_2 + \pi_3 f_3, \pi_i \geq 0, \sum_{i=1}^{3} \pi_i = 1 \right\} $$

is the convex hull of $\{f_1, f_2, f_3\}$, the hatched set in figure 2.1.

2.2.2 Identifiability

Teicher (1963) revealed a sufficient condition for the class of mixtures of $m$ or fewer binomial distributions with the same number of support points $T+1$ to be identifiable is that $T > 2m - 1$. Here we formulate a sufficient condition under which the class $\mathcal{H}$ of mixtures (2.1) is identifiable, which partly extends the result of Teicher (1963).

**Theorem 2.1.** Let $f(\phi), \phi \in \Omega$ be a one parameter multinomial distribution family in the mixture (2.1), and let

$$ A(\phi_1, \cdots, \phi_{2m}) = \begin{pmatrix} f(0; \phi_1) & f(0; \phi_2) & \cdots & f(0; \phi_{2m}) \\ f(1; \phi_1) & f(1; \phi_2) & \cdots & f(1; \phi_{2m}) \\ \vdots & \vdots & \vdots & \vdots \\ f(T; \phi_1) & f(T; \phi_2) & \cdots & f(T; \phi_{2m}) \end{pmatrix}. $$

If

$$ \text{rank}\ A(\phi_1, \cdots, \phi_{2m}) = 2m $$

for all distinct $\phi_1, \cdots, \phi_{2m}$, then the class $\mathcal{H}$ of mixtures (2.1) is identifiable.
figure 2.1. The two-dimensional probability simplex $P^2$. 
Proof. Let

\[ f(Q_1) = \sum_{j=1}^{m} \pi_j f(\phi_j) \in \mathcal{H}, \]

\[ f(Q_2) = \sum_{j=1}^{m} \pi_j f(\phi_j) \in \mathcal{H}. \]

Suppose

\[ \sum_{j=1}^{m} \pi_j f(\phi_j) = \sum_{j=1}^{m} \pi_j f(\phi_j'), \] \hspace{1cm} (2.3)

then by rearranging (2.3) we obtain

\[ \sum_{j=1}^{l} d_j f(\phi_j) = 0, \] \hspace{1cm} (2.4)

where \( \phi_1, \cdots, \phi_l \) are the all distinct elements from

\[ \phi_1', \cdots, \phi_m', \phi_1'', \cdots, \phi_m''. \]

So \( m \leq l \leq 2m \) and \( d_j \) equals to some \( \pi_i' \) or \( \pi_i'' \) or \( \pi_i' - \pi_i'' \) depending on

\[ \phi_j \in \{\phi_1', \cdots, \phi_m'\} \setminus \{\phi_1'', \cdots, \phi_m''\} \]

or

\[ \phi_j \in \{\phi_1'', \cdots, \phi_m''\} \setminus \{\phi_1', \cdots, \phi_m'\} \]

or

\[ \phi_j \in \{\phi_1', \cdots, \phi_m'\} \cap \{\phi_1'', \cdots, \phi_m''\}. \]

The assumption of Theorem 2.1 implies

\[ \text{rank} \begin{pmatrix} f(0; \phi_1) & f(0; \phi_2) & \cdots & f(0; \phi_l) \\ f(1; \phi_1) & f(1; \phi_2) & \cdots & f(1; \phi_l) \\ \vdots & \vdots & \ddots & \vdots \\ f(T; \phi_1) & f(T; \phi_2) & \cdots & f(T; \phi_l) \end{pmatrix} = l. \]

Therefore the system of eqs.(2.4) has only a zero solution. Thus

\[ d_j = 0, \quad 1 \leq j \leq l, \]
which implies if a support point in one of the mixtures (2.3) is not common, then the corresponding weight is 0. Thus only support points that belong to both mixtures in (2.3) are left, and for any common support point \( \phi_j = \phi_j^1 = \phi_j^2 \),

\[
(\pi_j^1, \phi_j^1) = (\pi_j^2, \phi_j^2)
\]

since

\[
d_j = \pi_j^1 - \pi_j^2 = 0.
\]

Now we can conclude that \( Q_1 = Q_2 \) \( \Box \).

**Example 2.1.** Let

\[
F = \{ f(t; T, \theta), \ 0 \leq \theta \leq 1 \}
\]

be a one parameter family of binomial distributions, \( T \) being fixed. Teicher (1963) proved that a sufficient condition that the class \( \mathcal{H}_m \) of all mixtures of \( m \) or fewer elements in \( F \) be identifiable is that \( T \geq 2m - 1 \).

This sufficient condition is the same as that in Theorem 2.1 for this special case.

Here

\[
f(t; T, \theta) = \binom{T}{t} \theta^t (1 - \theta)^{T-t}, \quad t = 0, 1, \ldots, T.
\]

Thus for all distinct \( \theta_1, \theta_2, \ldots, \theta_{2m} \),

\[
A(\theta_1, \ldots, \theta_{2m}) = \begin{pmatrix}
f(0; T, \theta_1) & f(0; T, \theta_2) & \cdots & f(0; T, \theta_{2m}) \\
f(1; T, \theta_1) & f(1; T, \theta_2) & \cdots & f(1; T, \theta_{2m}) \\
\vdots & \vdots & \ddots & \vdots \\
f(T; T, \theta_1) & f(T; T, \theta_2) & \cdots & f(T; T, \theta_{2m})
\end{pmatrix}
\]

\[
= \begin{pmatrix}
(1 - \theta_1)^T & (1 - \theta_2)^T & \cdots & (1 - \theta_{2m})^T \\
\binom{T}{1} \theta_1(1 - \theta_1)^{T-1} & \binom{T}{1} \theta_2(1 - \theta_2)^{T-1} & \cdots & \binom{T}{1} \theta_{2m}(1 - \theta_{2m})^{T-1} \\
\vdots & \vdots & \ddots & \vdots \\
\theta_1^T & \theta_2^T & \cdots & \theta_{2m}^T
\end{pmatrix}
\]
Via elementary row operations, the above matrix can be transformed into a Vander-
monde matrix

\[
V(\theta_1, \cdots, \theta_{2m}) = \begin{pmatrix}
1 & 1 & \cdots & 1 \\
\theta_1 & \theta_2 & \cdots & \theta_{2m} \\
\theta_1^2 & \theta_2^2 & \cdots & \theta_{2m}^2 \\
\vdots & \vdots & \ddots & \vdots \\
\theta_1^T & \theta_2^T & \cdots & \theta_{2m}^T
\end{pmatrix}.
\]

Here

\[\text{rank} V(\theta_1, \cdots, \theta_{2m}) = 2m\]

if and only if \( T \geq 2m - 1 \). So

\[\text{rank} A(\theta_1, \cdots, \theta_{2m}) = 2m\]

if and only if \( T \geq 2m - 1 \). Therefore for the binomial mixtures, the sufficient condition
in Theorem 2.1 is exactly the condition given by Teicher (1963).

Lindsay (1995) analyzed the identifiability of multinomial mixtures (2.1) from a
geometric point of view. He gave the condition of identifiability under a Tchebycheff
system density. We introduce the concept of a Tchebycheff system according to Karlin
and Studden (1966).

**Definition 2.8.** Let \( u_0(\phi), \cdots, u_T(\phi) \) denote continuous real-valued functions defined
on a closed finite interval \([a, b]\). These functions will be called a Tchebycheff system
provided the determinants

\[
\begin{vmatrix}
u_0(\phi_0) & u_0(\phi_1) & \cdots & u_0(\phi_T) \\
u_1(\phi_0) & u_1(\phi_1) & \cdots & u_1(\phi_T) \\
\vdots & \vdots & \ddots & \vdots \\
u_T(\phi_0) & u_T(\phi_1) & \cdots & u_T(\phi_T)
\end{vmatrix}
\]

never vanish whenever \( a \leq \phi_0 < \cdots < \phi_T \leq b \).
Definition 2.9. For any continuous function $f$ on $[a, b]$, we shall call an isolated zero $\phi_0 \in (a, b)$ of $f$ a nonnodal zero provided the function $f$ does not change sign at $\phi_0$. All other zeros including those at $a$ or $b$ are called nodal zeros. The number of zeros of $f$, where nodal zeros are counted once and nonnodal zeros twice is designated by $Z(f)$.

In figure 2.2, the points $\phi_1$ and $\phi_3$ are nonnodal zeros while $\phi_2$ and the endpoint $a$ are nodal zeros.

Definition 2.10. Let the system $\{u_i\}_0^T$ be defined on $[a, b]$. A function of the form $u = \sum_{i=0}^T a_i u_i$ where $a_i$ are real numbers will be called a polynomial. A polynomial is said to be nontrivial if $\sum_{i=0}^T a_i^2 > 0$.

Lemma 2.2. $\{u_i\}_0^T$ is a Tchebycheff system if and only if for every nontrivial polynomial $u = \sum_{i=0}^T a_i u_i$, $Z(u) \leq T$.

Let $a \leq \phi_1 < \cdots < \phi_k \leq b$. If we are to construct a nonnegative polynomial that vanishes at each of $\phi_i$ then each interior point $\phi_i$ will necessarily be a nonnodal zero while if the endpoints $a = \phi_1$ or $b = \phi_k$, these points will be nodal zeros by definition. We therefore assign to each $\phi_i$ a weight $w(\phi_i)$ defined by

$$w(\phi_i) = \begin{cases} 2, & \phi_i \in (a, b); \\ 1, & \phi_i = a \text{ or } b. \end{cases}$$

Lemma 2.3. (Krein Theorem) If $\{u_i\}_0^T$ is a Tchebycheff system and $\sum_{i=1}^k w(\phi_i) \leq T$, then there exists a nontrivial nonnegative polynomial $u = \sum_{i=0}^T a_i u_i$ vanishing precisely at the points $\phi_i$. The only exception is that if $T = 2m$ and exactly one of the endpoints $a$ or $b$ is in the set $\{\phi_i\}_{i=1}^k$, then $u$ may vanish at the other endpoint as well.

Suppose $f(t; \phi)$, $t = 0, 1, \cdots, T; \phi \in \Omega$ is a family of multinomial densities, then we have the following theorem.
figure 2.2.
Theorem 2.2. In the mixture
\begin{equation}
\pi_1 f(t; \phi_1) + \cdots + \pi_m f(t; \phi_m), \quad t = 0, \cdots, T,
\end{equation}
where \( m \) distinct \( \phi_1, \cdots, \phi_m \) are known and \( f(0; \phi), f(1; \phi), \cdots, f(T; \phi) \) constitute a Tchebycheff system, the parameters \( \pi_1, \cdots, \pi_m \) are identifiable provided that \( m \leq T + 1 \).

Proof. We can write (2.5) in the vector form
\begin{equation}
\pi_1 f(\phi_1) + \cdots + \pi_m f(\phi_m).
\end{equation}
If \( \pi_1, \cdots, \pi_m \) are not identifiable when \( m \leq T + 1 \), then there exist \( T + 1 \) distinct \( \phi_1, \cdots, \phi_{T+1} \) (the first \( m \) of the \( \phi_i \) are known) such that the matrix \( [f(\phi_1), \cdots, f(\phi_{T+1})] \) is singular. So the system of equations
\begin{equation}
[f(\phi_1), \cdots, f(\phi_{T+1})]'a = 0
\end{equation}
has a nonzero solution \( a \). By writing out equations (2.7) row by row, we obtain
\[ \sum_{t=0}^{T} a_t f(t; \phi_i) = 0, \quad i = 1, \cdots, T + 1. \]
Therefore \( \sum_{t=0}^{T} a_t f(t; \phi) \) has \( T + 1 \) zeros \( \phi_1, \cdots, \phi_{T+1} \), which due to Lemma 2.2 contradicts the fact that \( f(0; \phi), f(1; \phi), \cdots, f(T; \phi) \) is a Tchebycheff system. Hence we derive the desired conclusion. □

In figure 2.3, we illustrate Theorem 2.2 using the binomial model \( \text{Bin}(T, \theta) \), \( T = 2, \ 0 \leq \theta \leq 1 \). For each value of \( \theta \), the binomial density vector is
\[ f(\theta) = ((1-\theta)^2, 2\theta(1-\theta), \theta^2)' . \]
When \( \theta \) runs from 0 to 1, we trace out a curve \( C \) in the probability simplex \( P^2 \),
\[ C = \{ f(\theta) : \theta \in [0,1] \} , \]
figure 2.3. The unicomponent binomial density curve $C$ in the probability simplex $P^2$. 
and the corresponding set of binomial mixtures is the convex hull of \( C \), denoted by \( M = \text{conv}(C) \). From figure 2.3, we can see that any three binomial density vectors \( f(\theta_1), f(\theta_2), f(\theta_3) \) are linearly independent, so in the mixture

\[
f(Q) = \pi_1 f(\theta_1) + \pi_2 f(\theta_2) + \pi_3 f(\theta_3),
\]

(2.8)

\( \pi_1, \pi_2, \pi_3 \) are identifiable. We can also reach the same conclusion using Theorem 2.2 without the assistance of figure 2.3. In the mixture (2.8), \( T = 2, m = 3 \), and \( f(t; \theta), t = 0, 1, 2 \) constitute a Tchebycheff system. Since \( m = T + 1 \), following Theorem 2.2 we know that \( \pi_1, \pi_2, \pi_3 \) are identifiable.

We now consider the mixtures for which the locations of the support points \( \phi \) of \( Q \) are unknown. We use figure 2.3 to analyze the identifiability of the mixing distribution \( Q \) in such models.

(i) If \( f(Q) \in C \), then \( m = 1 \). In this situation,

\[
f(Q) = f(\theta)
\]

for some \( \theta \in [0, 1] \), and there is no other representation for \( f(Q) \), since every point in \( C \) is an extreme point of the convex set \( M \). In this case, \( Q \) is identifiable.

(ii) If \( f(Q) \in (f(0), f(1)) \), then \( m = 2 \). Here

\[
f(Q) = \pi_1 f(0) + \pi_2 f(1)
\]

for a unique pair of \( \pi_1, \pi_2 \) with \( 0 < \pi_1, \pi_2 < 1 \) and \( \pi_1 + \pi_2 = 1 \). Thus \( Q \) is identifiable.

(iii) If \( f(Q) \in M^0 \), the interior of \( M \) relative to the probability simplex, then one can draw infinitely many lines passing through \( f(Q) \) and joining two points on \( C \). Hence \( Q \) is not identifiable.

If we denote the number of support points of \( Q \) in \((a, b)\) by \( n_1 \) and the number of support points of \( Q \) in \( \{a, b\} \) by \( n_2 \), then we define the index of \( Q \) by
Thus for the above binomial mixture, \( Q \) is identifiable if and only if \( \text{index}(Q) \leq 1 \).

Summarizing the binomial case, we can obtain a general result.

**Theorem 2.3.** If \( f(t; \theta), \ t = 0, 1, \cdots, T \) is a Tchebycheff system density family, \( \theta \in [a, b] \), then the following three statements are equivalent.

(a) The mixing distribution \( Q \) in \( f(Q) \) is identifiable.

(b) The mixture \( f(Q) \) is on the relative boundary of the mixture set \( M \).

(c) \( \text{index}(Q) \leq \frac{T}{2} \).

**Proof.** (b) ⇒ (c). If \( f(Q) \) is on the relative boundary of the mixture set \( M \), denoted by \( f(Q) \in \text{rbd}(M) \), then there exists a supporting hyperplane

\[
H(v) : a'v = 0
\]

such that for all \( v \in M, a'v \leq 0 \) and \( a'f(Q) = 0 \). If \( \theta_0 \) is a support point of \( Q \), then \( f(\theta_0) \in H(v) \), i.e., \( a'f(\theta_0) = 0 \). Otherwise we will have \( a'f(Q) < 0 \). By Lemma 2.2 and the Tchebycheff property of the system \( f(t; \theta), \ t = 0, 1, \cdots, T \),

\[
Z(a'f(\theta)) \leq T.
\]

Moreover, since \( a'f(\theta) \leq 0 \) for all \( \theta \), any root in \( (a, b) \) must be nonnodal zeros of \( a'f(\theta) \).

Suppose the number of support points of \( Q \) in \( (a, b) \) is \( n_1 \) and the number of support points of \( Q \) in \( \{a, b\} \) is \( n_2 \), then

\[
2n_1 + n_2 \leq T,
\]

which implies

\[
n_1 + \frac{1}{2}n_2 = \text{index}(Q) \leq \frac{T}{2}.
\]
(b)⇒(a). From the above argument, we know that for any possible support point \( \theta_0 \) of \( Q \), \( f(\theta_0) \) is on the supporting hyperplane \( a'v = 0 \), i.e., \( a'f(\theta_0) = 0 \). According to Lemma 2.2, the number \( m \) of all possible support points of \( Q \) satisfies

\[
m = n_1 + n_2 \leq 2n_1 + n_2 \leq Z(a'f(\theta)) \leq T,
\]

and following Theorem 2.2, \( Q \) is identifiable.

(c)⇒(b). Suppose \( \text{index}(Q) \leq \frac{T}{2} \) and

\[
\theta_1, \theta_2, \ldots, \theta_{n_1+n_2}
\]

are the all support points of \( Q \), the first \( n_1 \) of the \( \theta_i \)'s (if any) are in \( (a, b) \), while the rest are in \( \{a, b\} \), then

\[
\sum_{i=1}^{n_1+n_2} w(\theta_i) = 2n_1 + n_2 = 2\text{index}(Q) \leq T.
\]

According to Lemma 2.3, there exists a nontrivial polynomial

\[
u(\theta) = \sum_{i=0}^{T} a_i f(i, \theta) \leq 0
\]

for all \( \theta \in [a, b] \) such that

\[
u(\theta_i) = 0, \quad i = 1, \ldots, n_1 + n_2
\]

and \( \theta_1, \ldots, \theta_{n_1} \) are the all zeroes of \( u(\theta) \) in \( (a, b) \).

Let \( a = (a_0, \ldots, a_T)' \), then

\[
0 = u(Q) = \int_a^b u(\theta)dQ(\theta) = \int_a^b \sum_{i=0}^{T} a_i f(i, \theta)dQ(\theta) = \sum_{i=0}^{T} a_i f(i, Q) = a'f(Q)
\]
and \( a'v \leq 0 \) for all \( v \in M \) but \( a'v \neq 0 \) on \( M \) since \( u(\theta) \) is nontrivial. So \( a'v = 0 \) is the supporting hyperplane passing through \( f(Q) \) and does not overlap the probability simplex, which implies \( f(Q) \in \text{rbd}(M) \).

(a) \( \Rightarrow \) (b). If \( f(Q) \) is not in the relative boundary of \( M \), i.e., \( f(Q) \in \text{ri}(M) \), the relative interior of \( M \), then for any \( \theta \in [a, b] \), there is a line passing through \( f(\theta), f(Q) \) and intersecting \( \text{rbd}(M) \). Therefore there exists a convex representation for \( f(Q) \) with \( f(\theta) \) as a component. We may conclude that the mixing distribution \( Q \) in \( f(Q) \) is not identifiable. \( \Box \)

For a Tchebycheff system density family, Theorem 2.3 is obviously more precise than Theorem 2.1. This is due to the fact that in Theorem 2.3, we use \( \text{index}(Q) \) to characterize identifiability while in Theorem 2.1 we use the number of support points of \( Q \) to determine the identifiability. For example, in figure 2.3, according to Theorem 2.1, we only know that mixtures with the number \( m \) of support points satisfying \( 2m - 1 \leq T = 2 \) or \( m \leq \frac{3}{2} \), that is, mixtures on the unicomponent density curve \( C = \{ f(\theta) : \theta \in [0, 1] \} \) are identifiable. By Theorem 2.3, a mixture \( f(Q) \) is identifiable if and only if \( \text{index}(Q) \leq \frac{T}{2} = 1 \). Only the mixtures in the subset \( C \cup (f(0), f(1)) \) have index less than or equal to 1, so the identifiable set is \( C \cup (f(0), f(1)) \). Hence we know mixtures on \( (f(0), f(1)) \) are also identifiable. These are ignored by Theorem 2.1. Since mixtures on \( (f(0), f(1)) \) have two support points, the criterion in Theorem 2.1 cannot distinguish these mixtures from those in the relative interior of \( M \) which also have a convex representation of two support points.

However, an advantage of Theorem 2.1 is that it can be applied to component density families beyond Tchebycheff systems, and its sufficient condition is the best one as far as the number of support points is concerned.
2.3 Results on General Mixtures

Teicher (1961, 1963) made an important contribution to the theory of identifiability of mixtures. He obtained some results that are applicable to a large amount of mixtures. Here we briefly state his results.

A family of one parameter distribution functions \( \{F(x; \phi), \phi \in \Omega\} \) is called an additively closed family if for any \( \phi_1, \phi_2 \in \Omega \), the convolution

\[
F(x; \phi_1) \ast F(x; \phi_2) = F(x; \phi_1 + \phi_2).
\]

Teicher (1961) made use of the form

\[
\phi(t; \alpha) = [\phi(t; 1)]^\alpha
\]

of characteristic functions of an additively closed family

\[
F = \{F(x; \alpha), \alpha \in \Omega\},
\]

where \( \Omega \) is the additive Abelian semi-group of (positive) integers (or rationals or real numbers). He asserted the class of mixtures of an additively closed family on such Abelian semi-group \( \Omega \) is identifiable.

Teicher (1961) also gave the conditions for a class of scale parameter mixtures and a class of translation parameter mixtures to be identifiable. More precisely, he asserted

(i) If the Fourier transform of \( F(e^x) \) is not identically zero in some nondegenerate real interval, then the class of scale parameter mixtures

\[
F(x; Q) = \int_0^\infty F(\phi x)dQ(\phi),
\]

where \( F(x) \) satisfies \( F(0+) = 0 \), is identifiable.

(ii) If the Fourier transform of \( F(x) \) is not identically zero in some nondegenerate real interval, then the class of translation parameter mixtures

\[
F(x; Q) = \int_{-\infty}^\infty F(x - \phi)dQ(\phi)
\]
Suppose \( f(x; \phi), \ x \in \mathcal{X}, \ \phi \in \Omega \) is a density family, where \( \mathcal{X} \) and \( \Omega \) are real intervals, and \( f(x; \phi) \) is Borel measurable on \( \mathcal{X} \times \Omega \). Let \( C_0(\Omega) \) be the Banach space of continuous functions on \( \Omega \) that vanish at \( \infty \). The Banach space \( C_0(\Omega) \) is normed by
\[
\|f\| = \sup\{|f(x)| : x \in \Omega\},
\]
\( f \in C_0(\Omega) \). Blum and Susarla (1977) derived another general result on identifiability by using the Riesz representation theorem. We state their result as follows.

Let
\[
\{f(x; \phi) : x \in \mathcal{X}\} \subset C_0(\Omega).
\]
Then the class of mixtures of \( \{f(x; \phi) : \phi \in \Omega\} \) is identifiable if and only if \( \{f(x; \phi) : x \in \mathcal{X}\} \) generates \( C_0(\Omega) \) in the supremum norm.

Teicher (1963, Theorem 2) obtained an important result on identifiability of finite mixtures:

Let \( \mathcal{F} = \{F\} \) be a family of distribution functions with transforms \( \phi(t) \) defined on \( S_\phi \) (the domain of \( \phi \)) such that the mapping \( M : F \rightarrow \phi \) is linear and one to one. Suppose that there exists a total ordering (\( \prec \)) of \( \mathcal{F} \) such that \( F_1 \prec F_2 \) implies
(i) \( S_{\phi_1} \subset S_{\phi_2} \); and
(ii) the existence of some \( t_1 \in S_{\phi_1} \) such that
\[
\lim_{t \to t_1} \phi_2(t)/\phi_1(t) = 0.
\]
Then the class of all finite mixtures of \( \mathcal{F} \) is identifiable.

As a consequence, the classes of all finite mixtures of normal distributions and Gamma distributions are identifiable. From this result of Teicher (1963), Yokowitz and Spragins (1968) and Mohanty (1972) concluded that the classes of finite mixtures of exponential, Cauchy, negative binomial, Poisson and Laguerre distributions...
are identifiable. Al-Hussaini and Ahmad (1981) used moment generating functions, characteristic functions and Laplace transforms of $X$ in Theorem 2 of Teicher (1963), where $X$ is the random variable with the probability distribution $F$, to show that classes of finite mixtures of the following ten univariate distribution families are identifiable:

1. logarithmic series,
2. discrete rectangular,
3. rectangular,
4. first law of Laplace,
5. noncentral $\chi^2$,
6. logistic,
7. generalized logistic,
8. generalized hyperbolic-secant,
9. inverse Gaussian,
10. random walk.

By using the moment generating function of log $X$ as the required transform in Theorem 2.4 of Chandra (1977) which states Theorem 2 of Teicher (1963) with a weaker assumption, Ahmad (1988) further proved the classes of finite mixtures of the following five distribution families are identifiable.

1. Weibull,
2. lognormal,
3. Chi,
4. Pareto,
5. power function.
Chapter 3

Nonparametric Maximum Likelihood Estimates (NPMLEs) of Mixing Distributions

3.1 Introduction

There are numerous statisticians attacking the problem of maximum likelihood estimation of a mixing distribution. Among them are the pioneers Robbins (1950), Kiefer and Wolfowitz (1956), Simar (1976), Laird (1978) and Jewell (1982).

Simar (1976) derived the existence, uniqueness and weak consistency of the MLE for the compounding distribution in a compound poisson process. He maximized the log likelihood function on a compact convex set (mixture set) and he showed that the resulting NPMLE has finite support points. He also gave the bounds for the number of its support points. Jewell (1982) did the same work as Simar (1976) for exponential mixtures.

Laird (1978) extended the results of Simar (1976) to arbitrary mixing distributions. She proved that the NPMLE of a mixing distribution is self-consistent. She then used the property of self-consistency to show that, under certain conditions, the NPMLE
must be a step function with a finite number of steps. However, the conditions given by her are very complicated.

Basing on these pioneer work Lindsay (1981, 1983) generalized the theory of non-parametric maximum likelihood estimation in mixture models from the geometric point of view. He obtained some elegant and precise properties of the NPMLE for mixing distributions by taking advantage of the convexity of the mixture likelihood set and concavity of the log likelihood function. We will summarize Lindsay's (1981, 1983) basic theory in Section 3.2 and also provide detailed proofs for some of his results.

However, all involving algorithms known so far lead to an estimator that is only an approximation to the NPMLE. Lindsay (1995) concluded: ”In most problems, one cannot hope to compute the NPMLE exactly because so far there is no finite step algorithm that will attain the solution.”

Seeking the NPMLE is still a challenge to statisticians. We are not ambitious to settle such a forbidding issue, but we can increase the knowledge on the location of the maximum likelihood vector that is a point on the boundary of the convex hull of the unicomponent likelihood curve. The relevant results are presented in Section 3.4.

In Section 3.5, we provide a detailed proof for the general version of self-consistency of the NPMLE for a mixing distribution and establish the conditions for the generalized likelihood equations. We give an example to illustrate that Lindsay’s (1981) original conditions cannot rationalize his proof. Finally, Section 3.6 gives a brief discussion on consistency of the NPMLE in mixture models and geometric estimation in binomial mixtures.
3.2 Lindsay’s NPMLE Theory and Its Supplement

Suppose we have a family of density functions \( f(x; \phi), \phi \in \Omega \). Let \( x_1, x_2, \ldots, x_n \) be a sample from the mixture distribution

\[
f(x; Q) = \int_\Omega f(x; \phi) dQ(\phi),
\]

where \( Q \) is a probability measure on \( \Omega \). Then the likelihood function is

\[
L(Q) = \prod_{i=1}^n L_i(Q) = \prod_{j=1}^D \left[ \int_\Omega L_j(\phi) dQ(\phi) \right]^{n_j},
\]

where \( L_j(\phi) = f(x_j; \phi) \) is the likelihood kernel, namely, the one component likelihood for a single observation \( x_j \), and \( n_j \) is the number of times \( x_j \) was observed in the sample. Our objective is to estimate \( Q \) by maximizing

\[
l(Q) = \Sigma_{i=1}^D n_i \ln L_i(Q)
\]

over all probability measures \( Q \) on \( \Omega \).

3.2.1 Introduction

From the log likelihood function (3.1), we know that the value of \( l(Q) \) depends on the mixture likelihood vector

\[
L(Q) = (L_1(Q), L_2(Q), \ldots, L_D(Q))'.
\]

We can convert the problem of maximizing \( l(Q) \) over all probability measures \( Q \) into the problem of maximizing

\[
l(L(Q)) = l(Q)
\]

over all \( D \)-dimensional vectors \( L(Q) \), as \( Q \) runs over all probability measures on \( \Omega \). The benefit of doing so is that we can take advantage of the geometric properties of
the convex set

\[ \mathcal{M} = \left\{ \mathbf{L}(Q) = (L_1(Q), \cdots, L_D(Q))' : L_i(Q) = \int_{\Omega} L_i(\phi)dQ(\phi), Q \in \mathcal{P} \right\}, \]

where \( \mathcal{P} \) is the set of all probability measures on \((\Omega, \mathcal{B})\), with \( \mathcal{B} \) being a \( \sigma \)-algebra on \( \Omega \). If we define the likelihood curve \( \Gamma \) in the nonnegative orthant

\[ R_+^D = \{ \mathbf{p} = (p_1, \cdots, p_D)': p_i \geq 0 \} \]

as

\[ \Gamma = \{ \mathbf{L}(\phi) = (L_1(\phi), \cdots, L_D(\phi))' : \phi \in \Omega \}, \]

then \( \mathcal{M} \) is included in the closure of the convex hull of \( \Gamma \), denoted by \( \text{conv}(\Gamma) \). Now our reformulated problem becomes maximizing the concave function \( l(\mathbf{p}) \) over the convex set \( \mathcal{M} \subseteq \text{conv}(\Gamma) \), which can be resolved by applying numerical optimization theory.

Before we give the main results on the mixture NPMLE, we take a brief study on \( l(\mathbf{p}) \). Specifically, \( l(\mathbf{p}) \) is a function from \( R_+^D \) to \( \mathbb{R} \), precisely

\[ l(\mathbf{p}) = \begin{cases} \sum_{i=1}^{D} n_i \ln(p_i), & p_i > 0 \text{ for all } i; \\ -\infty, & p_i = 0 \text{ for some } i. \end{cases} \]

Let \( \mathbf{p}_0, \mathbf{p}_1 \) be two points in \( R_+^D \). If we define

\[ \mathbf{p}_\varepsilon := (1 - \varepsilon)\mathbf{p}_0 + \varepsilon\mathbf{p}_1, \]

then \( \mathbf{p}_\varepsilon \) traces a linear path from \( \mathbf{p}_0 \) to \( \mathbf{p}_1 \), when \( \varepsilon \) goes from 0 to 1. We view

\[ l(\mathbf{p}_\varepsilon) = \sum_{i=1}^{D} n_i \ln((1 - \varepsilon)p_{0i} + \varepsilon p_{1i}) \]

as a function of \( \varepsilon \) and differentiate it with respect to \( \varepsilon \):

\[ l'(\mathbf{p}_0, \mathbf{p}_1) := \frac{d}{d\varepsilon} l((1 - \varepsilon)\mathbf{p}_0 + \varepsilon\mathbf{p}_1) \big|_{\varepsilon=0} = \sum_{i=1}^{D} n_i \left( \frac{p_{1i}}{p_{0i}} - 1 \right), \quad (3.2) \]
which is referred to as the Gateaux derivative of $l$ at $p_0$ to $p_1$. This Gateaux derivative is different from the directional derivative of $l$ at $p_0$ in direction $p_1 - p_0$ by a constant factor. Note that

$$\frac{d^2}{d\varepsilon^2} l((1 - \varepsilon)p_0 + \varepsilon p_1) = -\sum_{i=1}^{D} n_i \frac{(p_{1i} - p_{0i})^2}{p_{0i}^2} < 0 \quad \text{if} \quad p_0 \neq p_1,$$

which means $l(p)$ is strictly concave along any line segment $(p_0, p_1)$. An extremely important consequence of (3.3) is

$$l(p_1) < l(p_0) + l'(p_0, p_1) \quad \text{for any} \quad p_0 \neq p_1,$$

which is called the likelihood-Gateaux derivative inequality.

Because the gradient of $l$ at $p_0$ is the vector

$$\nabla l(p_0) = \left(\frac{n_1}{p_{01}}, \ldots, \frac{n_D}{p_{0D}}\right),$$

the directional derivative of $l$ at $p_0$ in direction $p_1 - p_0$ is

$$\nabla l(p_0) \cdot \frac{p_1 - p_0}{\|p_1 - p_0\|} = \frac{\sum_{i=1}^{D} n_i (p_{1i}/p_{0i}) - 1}{\|p_1 - p_0\|}.$$

Since $l(p)$ is strictly concave along $(p_0, p_1)$,

$$\frac{l(p_1) - l(p_0)}{\|p_1 - p_0\|} < \frac{\sum_{i=1}^{D} n_i (p_{1i}/p_{0i}) - 1}{\|p_1 - p_0\|}.$$

After simplification, we have

$$l(p_1) < l(p_0) + l'(p_0, p_1).$$

### 3.2.2 Fundamental Mixture NPMLE Theorems

The following theorem asserts the existence and uniqueness of maximizer $\hat{L}$ of $l$ over $\mathcal{M}$. We always assume that $\mathcal{M}$ contains at least one point with positive likelihood.
Theorem 3.1. Suppose that $\Gamma$ is closed and bounded. Then there exists a unique maximizer $\hat{L}$ of $l(p)$ over $M$, and $\hat{L} \in \partial M$, the boundary of $M$.

In Lindsay's (1983) paper, no detailed proof is given for this theorem.

Proof. Since $l(p)$ is a continuous concave function, the upper set

$$U_c = \{p \in \mathbb{R}_+^D : l(p) \geq c\}$$

is a closed convex set. Since $\Gamma$ is closed and bounded, so is $M$. The log likelihood $l$ takes its maximum value $c_0$ at some point $\hat{L}$ in $M$. We know for all $p \in M$,

$$l(p) \leq c_0$$

and $\hat{L} \in U_{c_0}$. We will show that $\hat{L} \in \partial M$.

If $\hat{L} \in M^0$, the interior of $M$, then there is an $\varepsilon > 0$ such that the open ball $O(\hat{L}, \varepsilon)$ centered at $\hat{L}$ with radius $\varepsilon$ is contained in $M$. Thus there are two cases:

(i). $[O(\hat{L}, \varepsilon) \setminus \hat{L}] \cap U_{c_0} = \phi$.

Let $p_1 \neq \hat{L}$ and $p_1 \in U_{c_0}$. Since $U_{c_0}$ is convex,

$$p_\pi = (1 - \pi)\hat{L} + \pi p_1 \in U_{c_0}$$

for any $0 \leq \pi \leq 1$. When $\pi$ is a small enough positive number, $p_\pi$ is in $O(\hat{L}, \varepsilon) \setminus \hat{L}$. So (i) cannot be true.

(ii). $[O(\hat{L}, \varepsilon) \setminus \hat{L}] \cap U_{c_0} \neq \phi$.

Suppose

$$p_2 \in [O(\hat{L}, \varepsilon) \setminus \hat{L}] \cap U_{c_0}.$$

We must have

$$l(p_2) = c_0.$$
Take a \( \pi \in (0,1) \), then
\[
(1 - \pi) \hat{L} + \pi p_2 \in \mathcal{M}. \tag{3.5}
\]

On the other hand, strict concavity of \( l \) implies
\[
l((1 - \pi) \hat{L} + \pi p_2) > c_0,
\]
which contradicts (3.5). Thus (ii) also cannot be true, and therefore we conclude \( \hat{L} \in \partial \mathcal{M} \). The same argument in discussion of case (ii) also gives the proof of the uniqueness of maximizer \( \hat{L} \).

From the above proof, we know that if \( \mathcal{M} \) is compact, then the conclusion of Theorem 3.1 holds. In van der Vaart and Wellner (1992), a sufficient condition is presented for \( \mathcal{M} \) to be compact:

**Proposition 3.1.** Let \( \Omega \) be a locally compact separable metric space with Borel \( \sigma \)-algebra \( \mathcal{B} \). If for each fixed \( x \) the function \( f(x; \phi) \in C_0(\Omega) \), the set of all continuous functions vanishing at infinity, then \( \mathcal{M} \) is a compact set.

The function \( f(x; \phi) \) vanishing at infinity means that for every \( \varepsilon > 0 \), there is a compact set \( C \subset \Omega \) such that
\[
|f(x; \phi)| < \varepsilon
\]
if \( \phi \notin C \). The families of Poisson distributions, binomial distributions, exponential distributions and mean-parameterized normal distributions satisfy the condition in Proposition 3.1, so for the mixtures of the above families, Theorem 3.1 holds. Hereinafter, we use \( \hat{Q} \) to signify the mixing distribution corresponding to \( \hat{L} \), which is called the NPMLE of \( Q \), i.e., \( \hat{L} = L(\hat{Q}) \). The next theorem gives a characterization of \( \hat{L} \).

**Theorem 3.2.** The following three statements are equivalent:
(a) \( L(\hat{Q}) \) maximizes \( l(L(Q)) \).

(b) \( L(\hat{Q}) \) minimizes \( \sup_{\phi} l'(L(Q), L(\phi)) \).

(c) \( \sup_{\phi} l'(L(\hat{Q}), L(\phi)) = 0 \).

Theorem 3.3 and Theorem 3.4 describe the properties of the support of an NPMLE \( \hat{Q} \).

**Theorem 3.3.** The support of any maximum likelihood estimator \( \hat{Q} \) lies in the set

\[
\{ \phi : l'(L(\hat{Q}), L(\phi)) = 0 \}.
\]

Theorem 3.1, Theorem 3.2 and Theorem 3.3 constitute a geometric image about \( \mathcal{M} \), the upper set \( U_\alpha \) and the support of \( \hat{Q} \). Following Theorem 3.3, for any support point \( \phi \) of a maximum likelihood estimator \( \hat{Q} \), the vector \( L(\phi) \) is on the hyperplane

\[
H := \{ p : l'(L(\hat{Q}), p) = 0 \}
\]

which obviously contains the point \( \hat{L} \). Theorem 3.2 specifies that this hyperplane \( H \) is a support hyperplane to the convex set \( \mathcal{M} \) at \( \hat{L} \), since by statement (c) and linearity of \( l'(\hat{L}, p) \) in \( p \),

\[
l'(\hat{L}, L(Q)) \leq 0,
\]

for all \( Q \in \mathcal{P} \). From the likelihood-Gateaux derivative inequality,

\[
l'(\hat{L}, p) > 0
\]

for all \( p \in U_\alpha \setminus \hat{L} \), which means \( U_\alpha \) is on the other side of \( H \). The maximizer \( \hat{L} \) is the only contact point of two convex sets \( \mathcal{M} \) and \( U_\alpha \). See figure 3.1. Application of Caratheodory's theorem gives further description of \( \hat{Q} \).
figure 3.1. The graph about the convex set $\mathcal{M} = \text{conv}(\Gamma)$, the upper set $U_{co}$ and the support hyperplane $H$. 
**Theorem 3.4.** There exists a mixture NPMLE $\hat{Q}$ which has no more than $D$ support points.

By Theorem 3.4, the MLE $\hat{Q}$ lies in the subset $\mathcal{P}_D$ of the set $\mathcal{P}$ of all probability distribution functions on $\Omega$, with support points no more than $D$. So when the sample is chosen, the problem of maximizing the likelihood over $\mathcal{P}$ is reduced to the problem of maximizing the likelihood over $\mathcal{P}_D$. Thus the support points and the weights of

$$\hat{Q}(\phi) = \sum_{j=1}^{D} \pi_j 1_{\{\phi_j\}}(\phi)$$

are the solution of the likelihood equations

$$\sum_{i=1}^{D} n_i \frac{\pi_j \partial f(x_i; \phi_j)}{\sum_{k=1}^{D} \pi_k f(x_i; \phi_k)} = 0, \quad j = 1, \ldots, D;$$

$$\sum_{i=1}^{D} n_i \frac{f(x_i; \phi_j) - f(x_i; \phi_D)}{\sum_{k=1}^{D} \pi_k f(x_i; \phi_k)} = 0, \quad j = 1, \ldots, D - 1 \quad (3.6)$$

subject to the restrictions $\pi_j \geq 0$ and $\sum_{j=1}^{D} \pi_j = 1$.

**3.2.3 Duality Method of Solving for $\hat{L}$**

Due to the complexity of the convex set $\mathcal{M}$, we change our maximization problem of maximizing $l(p)$ over $\mathcal{M}$ into

Problem 1: Minimize $l(p)$ in $R_+^D$ subject to the constraints

$$l'(p, L(\phi)) \leq 0,$$

for all $\phi \in \Omega$, or

Problem 2: Maximize $l(w)$ in $R_+^D$ subject to the constraints

$$\sum_{i=1}^{D} w_i L_i(\phi) \leq n,$$
for all $\phi \in \Omega$.

Problems 1 and 2 are equivalent. Since

$$l'(p, L(\phi)) = \sum_{i=1}^{D} n_i \left( \frac{L_i(\phi)}{p_i} - 1 \right) = \sum_{i=1}^{D} \frac{n_i}{p_i} L_i(\phi) - n,$$

let

$$w_i = \frac{n_i}{p_i},$$

then

$$l'(p, L(\phi)) \leq 0$$

if and only if

$$\sum_{i=1}^{D} w_i L_i(\phi) \leq n.$$

In addition,

$$l(w) = \sum_{i=1}^{D} n_i \ln w_i = \sum_{i=1}^{D} n_i \ln \frac{n_i}{p_i}$$

$$= \sum_{i=1}^{D} n_i \ln n_i - \sum_{i=1}^{D} n_i \ln p_i = \sum_{i=1}^{D} n_i \ln n_i - l(p). \quad (3.7)$$

Equality (3.7) shows that minimizing $l(p)$ is equivalent to maximizing $l(w)$. Thus, our primal maximization problem can be transformed into Problem 1 or 2 because of the following proposition.

**Proposition 3.2.** The upper set

$$U_{c_0} = \{ p : l(p) \geq c_0 = l(L) \}$$

includes the set

$$\{ p : l'(p, L(\phi)) \leq 0 \text{ for all } \phi \in \Omega \}.$$
Proof. If
\[ l(p) < l(\hat{L}), \]
then by the likelihood-Gateaux derivative inequality, we learn
\[ l'(p, \hat{L}) > 0, \]
where
\[ \hat{L} = \int_{\Omega} L(\phi) d\hat{Q}(\phi). \]
So
\[ l'(p, \hat{L}) = \int_{\Omega} l'(p, L(\phi)) d\hat{Q}(\phi) > 0, \]
there must exist \( \phi \in \Omega \) such that
\[ l'(p, L(\phi)) > 0, \]
which completes the proof. \( \square \)

We denote the set
\[ \{ p : l'(p, L(\phi)) \leq 0 \text{ for all } \phi \in \Omega \} \]
by \( A \). Since
\[ U_{\alpha} \supseteq A, \]
\[ l(p) \geq l(\hat{L}) \]
for all \( p \in A \). From Theorem 3.2, we also know \( \hat{L} \in A \), so \( \hat{L} \) is a minimizer of \( l \) over \( A \). If
\[ l(p') = l(\hat{L}) \]
and
\[ p' \neq \hat{L}, \]
then again the likelihood-Gateaux derivative inequality implies

\[ l'(p', \hat{L}) > 0. \]

By the previous argument, there exists \( \phi \in \Omega \) such that

\[ l'(p', L(\phi)) > 0, \]

so \( p' \not\in A \). Therefore we conclude \( \hat{L} \) is the only minimizer of \( l \) over \( A \). Hence we have the following theorem.

**Theorem 3.5.** \( \hat{L} \) is the maximizer of \( l \) over \( \mathcal{M} \) if and only if \( \hat{L} \) solves Problem 1 or

\[ \hat{w} = (\hat{w}_1, \ldots, \hat{w}_D)' \]

with

\[ \hat{w}_i = \frac{n_i}{\hat{L}_i} \]

solves Problem 2.

### 3.3 Convex Geometry

In order to well understand the structure of an NPMLE and to find potential solving methods, it is helpful to study some related convex geometry. All points and sets that we discuss are in Euclidean space \( \mathbb{R}^D \).

**Definition 3.1.** A point \( z \) of a set \( M \) is called an extreme point of \( M \) if it belongs to no open interval \( (x, y) \subset M \).

If \( M \) is convex, this means that \( z \) never lies between two points of \( M \). We denote the set of all extreme points of \( M \) by \( Ep(M) \). The concept of an extreme point of a convex set is extremely important because of the following fundamental representation theorem.
Lemma 3.1. (Krein-Milman Theorem) A closed bounded convex set $M$ is the convex hull of the set $Ep(M)$ of its extreme points.

Observe that the set $Ep(M)$ in the theorem is minimal in the sense that if $C$ is any set of points such that $M = \text{conv}(C)$, then $C \supset Ep(M)$. (Rockafellar, 1970, Corollary 18.3.1). The set $Ep(M)$ of extreme points of a closed bounded convex set $M$ need not be closed. For example, let $C_1$ be a closed circular disk in $\mathbb{R}^3$, and let $C_2$ be a line segment perpendicular to $C_1$ whose midpoint is a relative boundary point of $C_1$. The convex hull $C$ of $C_1 \cup C_2$ is closed. However, the set of extreme points of $C$ consists of the two endpoints of $C_2$ and all the relative boundary points of $C_1$ other than the midpoint of $C_2$, and this set is not closed. See figure 3.2.

Applying the Krein-Milman Theorem to mixture models, if $\Gamma$ is compact, then

$$M = \text{conv}(\Gamma) = \text{conv}(Ep(M)),$$

so

$$Ep(M) \subseteq \Gamma$$

and for every support point $\phi$ of the MLE $\hat{Q}$,

$$L(\phi) \in Ep(M)$$

as long as $\hat{Q}$ is unique.

Definition 3.2. A point $x_0$ of a set $M$ is called an exposed point of $M$ if $x_0$ is a support point of a supporting hyperplane $H$ of $M$, for which $H \cap M = \{x_0\}$. We denote the set of all exposed points of $M$ by $Ep^0(M)$.

Lemma 3.2. (Straszewicz Theorem) For any closed convex set $M$, the set $Ep^0(M)$ of exposed points of $M$ is a dense subset of the set $Ep(M)$ of extreme points of $M$. 
Figure 3.2. $E_p(C)$ is the relative boundary of $C_1$ except the midpoint $p_0$ of $C_2$, plus the two endpoints of $C_2$. 
Thus every extreme point is the limit of some sequence of exposed points.

Convexity is a global property of a set. Whether a point is an extreme point of a given set cannot be determined locally, it depends on the whole set. This fact makes the identification of extreme points of a convex set generally not easy. However, identifying exposed points of a convex set is much easier. Theorem 3.6 gives a way to identify $E_p^0(M)$.

Let

$$S = \{x = (x_1, \ldots, x_D)' : \|x\|^2 = x_1^2 + \cdots + x_D^2 = 1\}$$

be the unit sphere in $R^D$. For any $x \in S$, let $L_x$ be the set of elements $L(\phi)$ in $\Gamma$ which maximize

$$L(\phi) \cdot x = \sum_{i=1}^D L_i(\phi)x_i$$

among all $\phi$ in $\Omega$.

**Theorem 3.6.** Denote the set $\{L(\phi) : \exists x \in S \text{ such that } L_x = \{L(\phi)\}\}$ by $P(\Gamma)$. Then $E_p^0(M) = P(\Gamma)$.

**Proof.** From the definition of the exposed point of $M$, if

$$L(\phi) \in E_p^0(M),$$

then there is a supporting hyperplane $H$ of $M$ passing through $L(\phi)$ and

$$H \cap M = \{L(\phi)\}.$$ 

Let $x \in S, x \perp H$. If the direction of $x$ is on the other side of $H$ against $M$, then $L(\phi)$ is the only element in $M$ which maximizes $L(\phi) \cdot x$ among all $\phi$ in $\Omega$. So $L(\phi) \in P(\Gamma)$. Conversely, if $L(\phi) \in P(\Gamma)$, then $\exists x \in S$ such that

$$L_x = \{L(\phi)\}.$$
Let $H$ be the hyperplane passing through $L(\phi)$ and $x \perp H$, then for any $y \in H$,

$$(y - L(\phi)) \cdot x = 0.$$ 

Since for any $\phi' \in \Omega$,

$$(L(\phi') - L(\phi)) \cdot x \leq 0,$$  

(3.8)

consequently, for any $y \in M$,

$$(y - L(\phi)) \cdot x \leq 0,$$

Thus $H$ is a supporting hyperplane of $M$, and following the definition of $P(\Gamma)$, equality holds in inequality (3.8) if and only if

$$L(\phi') = L(\phi),$$

so

$$H \cap M = \{L(\phi)\}.$$ 

Therefore $L(\phi) \in Ep^0(M)$ \hfill \Box.

Now we give an example to illustrate how to apply Theorem 3.6 to determine all exposed points of a convex set.

**Example 3.1.** Figure 3.3 is the unicomponent likelihood curve $\Gamma$ of the uniform distribution on $(0, \phi)$ when there are two observations, $x_1 = 1$ and $x_2 = 4$. There are three exposed points $(0, 0), (1, 0)$ and $(\frac{1}{4}, \frac{1}{4})$ of

$$M = \text{conv}(\Gamma).$$

Now let us determine $P(\Gamma)$.

$$\Gamma = \{L(\phi) : \phi > 0\},$$
Figure 3.3. The unicomponent likelihood curve $\Gamma$ of the uniform distribution on $(0, \phi)$. 
where
\[
L(\phi) = \begin{cases} 
(0,0), & \phi < 1, \\
(\frac{1}{\phi},0), & 1 \leq \phi < 4, \\
(\frac{1}{4},0) \text{ and } (\frac{1}{4},\frac{1}{4}), & \phi = 4, \\
(\frac{1}{\phi},\frac{1}{\phi}), & \phi > 4.
\end{cases}
\]

Here in order that \( \Gamma \) is closed, \( L(\phi) \) takes two values \((\frac{1}{4},0)\) and \((\frac{1}{4},\frac{1}{4})\) at \( \phi = 4 \).

For \( x = (x_1,x_2) \in S \),
\[
L(\phi) \cdot x = \begin{cases} 
0, & \phi < 1, \\
\frac{x_1}{\phi}, & 1 \leq \phi < 4, \\
\frac{x_1}{\phi} \text{ and } \frac{x_1}{\phi} + \frac{x_2}{\phi}, & \phi = 4, \\
\frac{x_1}{\phi} + \frac{x_2}{\phi}, & \phi > 4.
\end{cases}
\]

Clearly
\[
P(\Gamma) = \{(0,0), (\frac{1}{4}, \frac{1}{4}), (1,0)\}
\]
is exactly the set of exposed points of \( \text{conv}(\Gamma) \).

**Remark 3.1.** If \( \Gamma \) is compact, then we have the following relationship
\[
\Gamma \supseteq \text{cl}(Ep^0(\mathcal{M})) \supseteq Ep(\mathcal{M}) \supseteq Ep^0(\mathcal{M}) = P(\Gamma).
\]

We only need to show \( Ep(\mathcal{M}) \supseteq P(\Gamma) \).

Suppose \( L(\phi) \in P(\Gamma) \). If \( L(\phi) \in (y_1,y_2) \subseteq \mathcal{M} \), then for any \( x \in S \),
\[
L(\phi) \cdot x \leq \max\{y_1 \cdot x, y_2 \cdot x\}.
\] (3.9)
If we take $x$ as the vector in $S$ such that
\[ \{L(\phi)\} = L_x, \]
then inequality (3.9) reduces to an equality. Without loss of generality, we assume
\[ y_1 \cdot x = L(\phi) \cdot x. \]

By the Caratheodory's theorem,
\[ y_1 = \sum_{i=1}^{D+1} \pi_i L(\phi_i), \quad \pi_i \geq 0, \quad \sum \pi_i = 1. \]

Thus we conclude
\[ L(\phi_i) = L(\phi), \quad i = 1, \ldots, D + 1, \]
so
\[ y_1 = L(\phi), \]
which contradicts $L(\phi) \in (y_1, y_2)$. Therefore any element in $P(\Gamma)$ belongs to no open interval $(y_1, y_2) \subseteq \mathcal{M}$, that is $P(\Gamma) \subseteq Ep(\mathcal{M})$.

Due to Krein-Milman Theorem and Straszewicz Theorem, the Gateaux derivative characterization of a mixture NPMLE can be summarized by the following theorem.

**Theorem 3.7.** $Q$ is a mixture NPMLE if
\[ II'(L(\hat{Q}), L(\phi)) \leq 0 \quad \text{for any} \quad L(\phi) \in P(\Gamma). \quad (3.10) \]

In Lindsay's (1983) original theorem, inequality (3.10) should hold for any $L(\phi) \in \Gamma$. According to Theorem 3.7, given a candidate MLE $\hat{Q}$, we only need to verify inequality (3.10) for $L(\phi) \in P(\Gamma)$. 
3.4 Location of the Maximum Likelihood Vector \( \hat{L} \)

Lindsay (1983) solved \( \hat{Q} \) by putting the problem of likelihood maximization into the formal setting of numerical optimization theory. That is, we view the problem as:

maximize the objective function \( l(p) \) over \( \mathcal{M} \). However, there are many difficulties in deriving the maximizer

\[
\hat{L} = L(\hat{Q})
\]

because of complexity of the convex set \( \mathcal{M} \). Even if we change the problem into its dual, difficulties still remain, since we probably have infinitely many linear constraints. In this section, we give a way to express the exact location of a maximum likelihood vector \( \hat{L} \). This result may lead to some algorithms to search the maximum likelihood vector \( \hat{L} \).

We assume that \( \Gamma \) is compact. For the compactness of \( \Gamma \), sometimes we permit that \( L(\phi) \) has multiple values for certain \( \phi \) in \( \Omega \). The following lemma is of remarkable importance to our result.

**Lemma 3.3.** If \( x \) is a point in a convex set \( M \) of \( R^D \) where a concave function \( f \) has a finite local maximum over \( M \), then \( x \) is a point where \( f \) has its global maximum over \( M \).

This is one of the most striking consequences of convexity. The gradient of \( l \) at \( \hat{L} \) is the vector

\[
\nabla l(\hat{L}) = \left( \frac{n_1}{\hat{L}_1}, \cdots, \frac{n_D}{\hat{L}_D} \right).
\]

Since the Gateaux derivative of \( l \) at \( \hat{L} \) to the point \( p \) is the inner product of \( \nabla l(\hat{L}) \) and \( p - \hat{L} \), that is,

\[
l'(\hat{L}, p) = \langle \nabla l(\hat{L}), p - \hat{L} \rangle,
\]
if we define
\[ H_{p_0}(p) := \langle \nabla l(p_0), p - p_0 \rangle, \]
then the set
\[ \{ p : l'(\hat{L}, p) \geq 0 \} \]
is the half space
\[ H_{\hat{L}}(p) \geq 0, \]
including the point \( \hat{L} + \nabla l(\hat{L}) \) and with boundary hyperplane
\[ H_{\hat{L}}(p) = 0 \]
passing through \( \hat{L} \) and perpendicular to \( \nabla l(\hat{L}) \), see the hatched half space in figure 3.4.

The maximizer \( \hat{L} \) is a local maximum point of \( l \) over \( \mathcal{M} \), so the Gateaux derivative of \( l \) at \( \hat{L} \) to any point on \( \mathcal{M} \) is less than or equal to 0, i.e.,
\[ H_{\hat{L}}(p) \leq 0, \quad \forall p \in \mathcal{M}, \]
it will suffice that
\[ H_{\hat{L}}(p) \leq 0, \quad \forall p \in P(\Gamma). \]
This is exactly the Gateaux derivative characterization of a mixture MLE. We now clearly see that the convex set \( \mathcal{M} \) is on the other side of the hyperplane
\[ H_{\hat{L}}(p) = 0, \]
so
\[ H_{\hat{L}}(p) = 0 \]
is the supporting hyperplane of \( \mathcal{M} \) at \( \hat{L} \). This is equivalent to saying
\[ \{ p : H_{\hat{L}}(p) > 0 \} \cap \mathcal{M} = \phi. \]
figure 3.4. The uniform mixture set $\mathcal{M}$, and the closed half space $H_L(p) \geq 0$. 
According as Lemma 3.3, any other point \( p_0 \neq \hat{L} \) in \( \mathcal{M} \) is not a local maximum point of \( l \) over \( \mathcal{M} \), so

\[
\{p : H_{p_0}(p) > 0\} \cap \mathcal{M} \neq \emptyset,
\]

as shown in figure 3.5. The hatched open half space

\[
\{p : H_{(1/2)}(p) > 0\}
\]

intersects \( \mathcal{M} \).

**Remark 3.2.** The gradient of the function \( l(p) \) at \( p \) is the vector

\[
\nabla l(p) = \left( \frac{n_1}{p_1}, \ldots, \frac{n_D}{p_D} \right).
\]

When some \( p_i = 0 \), we define \( \frac{n_i}{p_i} = \infty \). If all other \( p_j, j \neq i \) are nonzero, then we deem \( \nabla l(p) \) to be a vector perpendicular to the hyperplane \( p_i = 0 \). When

\[
p_{ij} = 0, \quad j = 1, \ldots, k,
\]

the \( p_{ij} \) are different elements in \( \{p_1, p_2, \ldots, p_D\} \), and for all other \( i \neq i_j, p_i \neq 0 \), \( \nabla l(p) \)

stands for a vector with direction \((0, \ldots, 0, n_{i_1}, 0, \ldots, 0, n_{i_2}, 0, \ldots, 0, n_{i_k}, 0, \ldots, 0)\), where the \( i_j \)th element is \( n_{i_j} \), that is

\[
\nabla l(p) \propto (0, \ldots, 0, n_{i_1}, 0, \ldots, 0, n_{i_2}, 0, \ldots, 0, n_{i_k}, 0, \ldots, 0).
\]

This vector is perpendicular to the subspace

\[
\{p : p_{ij} = 0, \quad j = 1, \ldots, k\}.
\]

In figure 3.5, the gradient of \( l \) at \((1, 0)\) is \( \nabla l((1,0)) \propto (0,1) \), hence the set

\[
\{p : (\nabla l((1,0)), p - (1,0)) > 0\}
\]
Figure 3.5. The open half spaces $H_{(\frac{1}{2}, \frac{1}{2})}(p) > 0$ and $H_{(1, 0)}(p) > 0$. 
is the dotted open half space \( \{H_{(1,0)}(p) > 0\} \). We also note that the maximum likelihood vector

\[
\hat{L} = \left( \frac{1}{2}, \frac{1}{6} \right) \in \{H_{(\frac{1}{4}, \frac{1}{4})}(p) > 0\} \cap \{H_{(1,0)}(p) > 0\} \cap M.
\]

When \( p_0 \in M \) and \( p_0 \neq \hat{L} \), we have

\[
l(p_0) < l(\hat{L}).
\]

Since \( l \) is a strictly concave function, the Gateaux derivative \( l'(p_0, \hat{L}) > 0 \), which means

\[
\hat{L} \in \{H_{p_0}(p) > 0\}.
\]

**Theorem 3.8 A.** The maximum likelihood vector \( \hat{L} \) is a unique element in \( M \) which belongs to the intersection of all open half space

\[
\{p : H_{p_0}(p) > 0\}, \quad p_0 \neq \hat{L}, \quad p_0 \in M,
\]

that is,

\[
\bigcap_{p_0 \in M} \{p : H_{p_0}(p) > 0\} \cap M = \{\hat{L}\}.
\]

**Proof.** We have already proven that \( \hat{L} \) lies in

\[
\bigcap_{p_0 \in M} \{p : H_{p_0}(p) > 0\} \cap M.
\]

Now \( \hat{L} \) is the only point in

\[
\bigcap_{p_0 \in M} \{p : H_{p_0}(p) > 0\} \cap M
\]

is obvious. Since, if

\[
p' \in M, \quad p' \neq \hat{L},
\]
then
\[ p' \notin \{ p : H_{p'}(p) > 0 \} \cap \mathcal{M}. \]

Therefore, \( p' \) is not in the set
\[ \bigcap_{p_0 \in \mathcal{M}} \{ p : H_{p_0}(p) > 0 \} \cap \mathcal{M}. \]

Hence \( \hat{L} \) is a unique element in
\[ \bigcap_{p_0 \in \mathcal{M}} \{ p : H_{p_0}(p) > 0 \} \cap \mathcal{M}. \]

We come to the conclusion that \( \hat{L} \) is a unique element in
\[ \bigcap_{p_0 \in \mathcal{M}} \{ p : H_{p_0}(p) > 0 \} \cap \mathcal{M}. \quad \square \]

We also can write the Theorem 3.8 A in a slightly different form:

**Theorem 3.8 B.** \( \bigcap_{p_0 \in \mathcal{M}} \{ p : H_{p_0}(p) \geq 0 \} \cap \mathcal{M} = \{ \hat{L} \}. \)

**Proof.** From Theorem 3.8 A,
\[ \hat{L} \in \bigcap_{p_0 \in \mathcal{M}} \{ p : H_{p_0}(p) \geq 0 \} \cap \{ p : H_{\hat{L}}(p) \geq 0 \} \cap \mathcal{M} \]
\[ = \bigcap_{p_0 \in \mathcal{M}} \{ p : H_{p_0}(p) \geq 0 \} \cap \mathcal{M}. \]

Suppose
\[ p' \in \mathcal{M}, \quad p' \neq \hat{L}. \]

Since \( p' \) is not local maximum point of \( l \) over \( \mathcal{M} \), there exists \( p'' \) in \( \mathcal{M} \) such that
\[ l(p'') > l(p'). \]
and the line segment \((p', p'')\) does not contain \(\tilde{L}\). Following the preceding argument,

\[
l'(p', p'') > 0.
\]

Let

\[
p_\varepsilon = (1 - \varepsilon)p' + \varepsilon p'',
\]

since

\[
\frac{d}{d\varepsilon} l(p_\varepsilon)|_{\varepsilon=0} = l'(p', p'') > 0
\]

and \(\frac{d}{d\varepsilon} l(p_\varepsilon)\) is a continuous function of \(\varepsilon\),

\[
\frac{d}{d\varepsilon} l(p_\varepsilon) = \sum_{i=1}^{D} n_i \frac{-p_i' + p_i''}{p_{\varepsilon i}} > 0, \text{ in } [0, \varepsilon_0), \text{ for some } 0 < \varepsilon_0 < 1. \quad (3.11)
\]

Taking \(\varepsilon = \frac{\varepsilon_0}{2}\) in (3.11), obviously

\[
p_{\varepsilon_0/2} \in \mathcal{M}, \quad p_{\varepsilon_0/2} \neq \tilde{L}
\]

and

\[
\frac{d}{d\varepsilon} l(p_{\varepsilon_0/2}) = \sum_{i=1}^{D} n_i \frac{-p_i' + p_i''}{p_{\varepsilon_0 i}} > 0.
\]

Thus

\[
l'(p_{\varepsilon_0/2}, p') = \frac{\varepsilon_0}{2} \cdot \sum_{i=1}^{D} n_i \frac{-p_i' + p_i''}{p_{\varepsilon_0 i}} = -\frac{\varepsilon_0}{2} \cdot \frac{d}{d\varepsilon} l(p_{\varepsilon_0/2}) < 0.
\]

So

\[
p' \notin \{p : H_{p_{\varepsilon_0/2}}(p) \geq 0\}.
\]

Hence,

\[
p' \notin \bigcap_{p_0 \in \mathcal{M}} \{p : H_{p_0}(p) \geq 0\} \cap \mathcal{M},
\]

which implies

\[
p' \notin \bigcap_{p_0 \in \mathcal{M}} \{p : H_{p_0}(p) \geq 0\} \cap \mathcal{M}.
\]
So \( L \) is the unique element in

\[
\cap_{p_0 \in M} \{ p : H_{p_0}(p) \geq 0 \} \cap M. \quad \square
\]

Indeed we have the following result.

**Theorem 3.8 C.** \( \cap_{p_0 \in \partial M} \{ p : H_{p_0}(p) \geq 0 \} \cap \partial M = \{ L \} \).

**Proof.** Evidently,

\[
L \in \cap_{p_0 \in \partial M} \{ p : H_{p_0}(p) \geq 0 \} \cap \partial M.
\]

Conversely, if

\[
p' \in \partial M \cap \{ p : H_L(p) = 0 \} = \partial M \cap \{ p : H_L(p) \geq 0 \}, \quad p' \neq L,
\]

then

\[
l(L) > l(p').
\]

The preceding argument indicates

\[
l'(p', L) > 0.
\]

Let

\[
p_\varepsilon = (1 - \varepsilon)p' + \varepsilon L.
\]

In similar way to that in the proof of Theorem 3.8 B, we can show that there is

\[
p_{\varepsilon_0} \in \partial M \cap \{ H_L(p) = 0 \}, \quad p_{\varepsilon_0} \neq L
\]

such that

\[
l'(p_{\varepsilon_0}, p') < 0
\]
for some $0 < \varepsilon_0 < 1$. Thus

$$p' \notin \{H_{p_0}(p) \geq 0\},$$

which implies

$$p' \notin \bigcap_{p_0 \in \partial M} \{H_{p_0}(p) \geq 0\} \cap \{H_L(p) = 0\} \cap \partial M$$

$$= \bigcap_{p_0 \in \partial M} \{H_{p_0}(p) \geq 0\} \cap \{H_L(p) = 0\} \cap \partial M$$

which completes the proof. \qed

Corresponding to Theorem 3.2 (c), we have the following corollary.

**Corollary 3.1.** If $\hat{Q}$ is MLE of $Q$, then

$$\inf_{\phi} I'(L(\phi), L(\hat{Q})) \geq 0.$$  

**Proof.**

$$I'(L(\phi), L(\hat{Q})) = \sum_{i=1}^{D} \frac{n_i}{f(x_i; \phi)} (\hat{L}_i - f(x_i; \phi))$$

$$= \left( \frac{n_1}{f(x_1; \phi)}, \ldots, \frac{n_D}{f(x_D; \phi)} \right) ((\hat{L}_1 - f(x_1; \phi)), \ldots, (\hat{L}_D - f(x_D; \phi)))$$

$$= H_{L_0}(\hat{L}) \geq 0$$

for all $\phi \in \Omega$ from Theorem 3.8 B. \qed

Finally we will describe a feature of the support hyperplane $H_L(p)$.

**Theorem 3.9.** If $x_0 \in \text{conv}(L_\infty)$ such that $\nabla I(x_0) \propto x$, then

$$Oth(x) = \{H_L(p) = 0\}$$
and

\[ x_0 = \hat{L}, \]

where \( \text{Oth}(x) \) stands for the hyperplane orthogonal to \( x \) and including \( L_x \).

**Proof.** Since \( x_0 \in \text{conv}(L_x), x_0 \in \partial \mathcal{M} \), by the definition of \( L_x \),

\[ \langle L(\phi) - x_0, x \rangle \leq 0, \text{ for any } \phi \in \Omega. \]

That is,

\[ \langle L(\phi) - x_0, \nabla l(x_0) \rangle \leq 0, \]

or

\[ l'(x_0, L(\phi)) \leq 0, \text{ for any } \phi \in \Omega. \]

By Theorem 3.2, \( x_0 = \hat{L} \). We also can conclude

\[ \text{Oth}(x) = \{ H_L(p) = 0 \}. \]

### 3.5 Self-Consistency of NPMLEs in Mixture Models

Laird (1978) showed the NPMLE for a mixing distribution is self-consistent, a property which suggests an iterative algorithm for computation. Lindsay (1981) gave a more general version of self-consistency of the NPMLE for a mixing distribution with fixed size of the support set. In the following theorem, we will state this result for an arbitrary mixing distribution.

If we treat the parameter \( \phi \) as a random variable with distribution function \( Q \), then the distribution of \( X \) is the marginal distribution of \( (X, \phi) \). The posterior expectation
of a function $g(\phi)$ is

$$E[g(\phi)|X = x; Q] = \frac{\int_{\Omega} g(\phi)f(x; \phi)dQ(\phi)}{\int_{\Omega} f(x; \phi)dQ(\phi)}.$$ 

In particular, when

$$g(\phi) = I\{\phi \leq \theta\},$$

we obtain the posterior distribution of $\phi$

$$P(\phi \leq \theta|X = x; Q) = \frac{\int_{\Omega} I\{\phi \leq \theta\}f(x; \phi)dQ(\phi)}{\int_{\Omega} f(x; \phi)dQ(\phi)}.$$ 

**Definition 3.3.** An estimate $\tilde{Q}$ of the distribution $Q$ based on the sample $x_1, \cdots, x_n$ from $f(x; Q)$ is called self-consistent if $\tilde{Q}$ satisfies

$$P(\phi \leq \theta; \tilde{Q}) := \int_{\Omega} I\{\phi \leq \theta\}d\tilde{Q}(\phi) = n^{-1}\sum_{i=1}^{n} P(\phi \leq \theta|X = x_i; \tilde{Q}).$$

**Theorem 3.10.** If $\hat{Q}$ is an NPMLE for the mixing distribution $Q$ and $g(\phi)$ is a function on $\Omega$, then

$$E[g(\phi); \hat{Q}] := \int_{\Omega} g(\phi)d\hat{Q}(\phi) = n^{-1}\sum_{i=1}^{n} E[g(\phi)|X = x_i; \hat{Q}]. \quad (3.12)$$ 

**Proof.** Let

$$dQ_t(\phi) = \frac{e^{t g(\phi)}d\hat{Q}(\phi)}{\int_{\Omega} e^{t g(\phi)}d\hat{Q}(\phi)},$$

then the log-likelihood under

$$f(x; Q_t) = \int_{\Omega} f(x; \phi)dQ_t(\phi)$$
\[ l = \sum_{i=1}^{n} \ln \int_{\Omega} f(x_i; \phi) dQ_t(\phi) \]
\[
= \sum_{i=1}^{n} \ln \int_{\Omega} f(x_i; \phi) \cdot \frac{e^{tg(\phi)} dQ(\phi)}{\int_{\Omega} e^{tg(\phi)} dQ(\phi)}
\]
\[
= \sum_{i=1}^{n} \ln \int_{\Omega} f(x_i; \phi) e^{tg(\phi)} dQ(\phi) - n \ln \int_{\Omega} e^{tg(\phi)} dQ(\phi).
\]

So
\[
\frac{dl}{dt} = \sum_{i=1}^{n} \frac{\int_{\Omega} g(\phi) f(x_i; \phi) e^{tg(\phi)} dQ(\phi)}{\int_{\Omega} f(x_i; \phi) e^{tg(\phi)} dQ(\phi)} - n \frac{\int_{\Omega} g(\phi) e^{tg(\phi)} dQ(\phi)}{\int_{\Omega} e^{tg(\phi)} dQ(\phi)}.
\]

Therefore, since \( \hat{Q} \) is a mixture NPMLE,
\[
\left. \frac{dl}{dt} \right|_{t=0} = 0,
\]

namely,
\[
\sum_{i=1}^{n} \frac{\int_{\Omega} g(\phi) f(x_i; \phi) dQ(\phi)}{\int_{\Omega} f(x_i; \phi) dQ(\phi)} - n \int_{\Omega} g(\phi) dQ(\phi) = 0,
\]
or
\[
n^{-1} \sum_{i=1}^{n} E[g(\phi)|X = x_i; \hat{Q}] = E[g(\phi); \hat{Q}]. \quad \Box
\]

In the following theorem, the condition is different from that in Lindsay's (1981) original theorem, suggesting that Lindsay's (1981) condition is not accurate. We should also notice that no matter what the support \( \Omega \) of \( Q \) is, \( Q \) is defined on the whole real axis \((-\infty, \infty)\). Thus, we can write all integrals on \( \Omega \) as the integrals on \((-\infty, \infty)\).

**Theorem 3.11.** Suppose that for any \( x \) in the sample space, \( \ln f(x; \phi) \) is differentiable with respect to \( \phi \) in \( \Omega \). Suppose further that \( g(\phi) \) is a differentiable function and for any \( t \) in a small neighborhood of 0, \( 1 + tg'(\phi) > \varepsilon \), for some small positive number \( \varepsilon \).
If $\hat{Q}$ is an NPMLE for the mixing distribution, then

$$\sum_{i=1}^{n} E \left[ g(\phi) \cdot \frac{\partial \ln f(X; \phi)}{\partial \phi} \bigg| X = x_i; \hat{Q} \right] = 0. \quad (3.13)$$

Moreover if $f(x; \phi)$ is twice differentiable in $\phi$ over $\Omega$, then

$$\sum_{i=1}^{n} E \left[ \frac{\partial^2 f(X; \phi)}{f(X; \phi)} \cdot g^2(\phi) - \frac{\partial \ln f(X; \phi)}{\partial \phi} \cdot g'(\phi)(1 - g(\phi)) \bigg| X = x_i; \hat{Q} \right]$$

$$\leq \sum_{i=1}^{n} \left\{ E \left[ g(\phi) \cdot \frac{\partial \ln f(X; \phi)}{\partial \phi} \bigg| X = x_i; \hat{Q} \right] \right\}^2.$$

**Proof.** Let

$$dQ_t(\phi) = d\hat{Q}(\phi + tg(\phi)),$$

it is easy to verify that $Q_t$ is a probability distribution function of $\phi$.

Consider the implicit function

$$F(\phi, t, \theta) = \phi + tg(\phi) - \theta = 0.$$

Since

$$\frac{\partial F(\phi, t, \theta)}{\partial \phi} = 1 + tg'(\phi) > \varepsilon > 0,$$

there is a unique implicit function $\phi = h(\theta, t)$ such that

$$F(h(\theta, t), t, \theta) = 0.$$
Thus $h(\theta, 0) = \theta$ and

$$\frac{\partial h(\theta, t)}{\partial t} = -\frac{\partial F/\partial t}{\partial F/\partial \phi}$$

$$= -\frac{g(\phi)}{1 + tg'(\phi)}$$

$$= -\frac{g(h(\theta, t))}{1 + tg'(h(\theta, t))}.$$  

We have already known $\hat{Q}$ has a finite number of support points. Assume that its support points are $\theta_1, \cdots, \theta_m$ with weights $\pi_1, \cdots, \pi_m$ respectively. The log-likelihood under $f(x; Q_t)$ is

$$l = \sum_{i=1}^{n} \ln \int_{-\infty}^{\infty} f(x_i; \phi) dQ_t(\phi)$$

$$= \sum_{i=1}^{n} \ln \int_{-\infty}^{\infty} f(x_i; \phi) d\hat{Q}(\phi + tg(\phi))$$

$$= \sum_{i=1}^{n} \ln \sum_{j=1}^{m} f(x_i; h(\theta_j, t))\pi_j.$$  

$$\frac{dl}{dt} = \sum_{i=1}^{n} \frac{\sum_{j=1}^{m} f_2(x_i; h(\theta_j, t)) \cdot \partial h(\theta_j, t)/\partial t \cdot \pi_j}{\sum_{j=1}^{m} f(x_i; h(\theta_j, t))\pi_j}$$

$$= \sum_{i=1}^{n} \frac{\sum_{j=1}^{m} f_2(x_i; h(\theta_j, t)) \left( -\frac{g(h(\theta_j, t))}{1 + tg'(h(\theta_j, t))} \right) \pi_j}{\sum_{j=1}^{m} f(x_i; h(\theta_j, t))\pi_j},$$

where

$$f_2(x; \phi) = \frac{\partial f(x; \phi)}{\partial \phi}.$$
\[
\frac{dl}{dt}\bigg|_{t=0} = \sum_{i=1}^{n} - \sum_{j=1}^{m} \frac{f_{2}(x_{i}; \theta_{j})g(\theta_{j})\pi_{j}}{\sum_{j=1}^{m} f(x_{i}; \theta_{j})\pi_{j}} \\
= - \sum_{i=1}^{n} \int_{\Omega} f_{2}(x_{i}; \phi)g(\phi)d\hat{Q}(\phi) \\
= - \sum_{i=1}^{n} \int_{\Omega} f(x_{i}; \phi)d\hat{Q}(\phi) \\
= - \sum_{i=1}^{n} E \left[ g(\phi) \cdot \frac{\partial \ln f(X; \phi)}{\partial \phi} \bigg| X = x_{i}; \hat{Q} \right]. \quad (3.14)
\]

Proceeding with a similar calculation, we obtain
\[
\frac{d^2l}{dt^2}\bigg|_{t=0} = \sum_{i=1}^{n} \left\{ \int_{\Omega} \left[ \frac{f_{22}(x_{i}; \phi)}{f(x_{i}; \phi)} g^2(\phi) \frac{\partial \ln f(X; \phi)}{\partial \phi} g'(\phi)(1 - g(\phi)) \right] f(x_{i}; \phi)d\hat{Q}(\phi) \\
- \left[ \frac{\int_{\Omega} \frac{\partial \ln f(x_{i}; \phi)}{\partial \phi} g(\phi)f(x_{i}; \phi)d\hat{Q}(\phi)}{\int_{\Omega} f(x_{i}; \phi)d\hat{Q}(\phi)} \right]^2 \right\} \\
= \sum_{i=1}^{n} E \left[ \frac{f_{22}(X; \phi)}{f(X; \phi)} g^2(\phi) - \frac{\partial \ln f(X; \phi)}{\partial \phi} g'(\phi)(1 - g(\phi)) \bigg| X = x_{i}; \hat{Q} \right] \\
- \sum_{i=1}^{n} \left\{ E \left[ \frac{\partial \ln f(X; \phi)}{\partial \phi} g(\phi) \bigg| X = x_{i}; \hat{Q} \right] \right\}^2, \quad (3.15)
\]

where
\[
f_{22}(x; \phi) = \frac{\partial^2 f(x; \phi)}{\partial \phi^2}.
\]

Since \( \hat{Q} \) maximizes \( l \),
\[
\frac{dl}{dt}\bigg|_{t=0} = 0 \quad \text{and} \quad \frac{d^2l}{dt^2}\bigg|_{t=0} \leq 0.
\]

By eq. (3.14) and (3.15),
\[
\sum_{i=1}^{n} E \left[ g(\phi) \cdot \frac{\partial \ln f(X; \phi)}{\partial \phi} \bigg| X = x_{i}; \hat{Q} \right] = 0
\]
\[
\sum_{i=1}^{n} E \left[ \frac{\partial^2 f(X; \phi)/\partial \phi^2}{f(X; \phi)} \cdot g^2(\phi) - \frac{\partial \ln f(X; \phi)}{\partial \phi} \cdot g'(\phi)(1 - g(\phi)) \right] \bigg| X = x_i; \hat{Q} \right]
\]
\[
\leq \sum_{i=1}^{n} \left\{ E \left[ g(\phi) \cdot \frac{\partial \ln f(X; \phi)}{\partial \phi} \bigg| X = x_i; \hat{Q} \right] \right\}^2. \quad \square
\]

**Remark 3.3.** Lindsay’s (1981) original condition: \( \Omega = [a, b] \) and \( g(\phi) = 0 \) on \( \partial \Omega \) cannot guarantee \( Q_t \) being a probability distribution function. For example, \( \Omega = [0, 1] \) and

\[
g(\phi) = \begin{cases} 
0, & \phi < \frac{1}{2}, \\
1, & \phi = \frac{1}{2}, \\
\frac{1}{2} \ln(2\phi - 1), & \frac{1}{2} < \phi < 1, \\
0, & \phi \geq 1
\end{cases}
\]

satisfies Lindsay’s condition.

\[
y(\phi) = \phi + tg(\phi) = \begin{cases} 
\phi, & \phi < \frac{1}{2}, \\
\phi + t, & \phi = \frac{1}{2}, \\
\phi + \frac{1}{2} \ln(2\phi - 1), & \frac{1}{2} < \phi < 1, \\
\phi, & \phi \geq 1.
\end{cases}
\]

See figure 3.6 for the features of the graph of \( y(\phi) = \phi + t g(\phi) \), where \( t < 0 \) with small absolute value \( |t| \). Clearly, in the interval \((-\infty, \frac{1}{2}]\),

\[
y(\phi) < \frac{1}{2};
\]

however, in the interval \((\frac{1}{2}, \infty)\),

\[
y(\phi) > \frac{1}{2} - \frac{t}{2}(1 - \ln(-t)) > \frac{1}{2}.
\]

Thus, \( y(\phi) \) does not take value \( \frac{1}{2} \) for any \( \phi \in R \).

If \( \hat{Q} \) has \( \theta = \frac{1}{2} \) as one of its support points with positive weight, then \( Q_t \) is not a probability distribution function for \( t \) in the left neighborhood of 0. Moreover,

\[
\int_{-\infty}^{\infty} dQ_t(\phi) = \int_{-\infty}^{\infty} d\hat{Q}(\phi + t g(\phi)) = \int_{-\infty}^{t+\frac{1}{2}} d\hat{Q}(\theta) < 1
\]
Figure 3.6. The graph of $y(t) = \phi + t\phi(t)$. 
if \( \hat{Q} \) has support points in \([\frac{1}{2}, \infty)\). We also note that \( Q_t(\phi) \) is not nondecreasing in \( \phi \) on \( R \).

Eq. (3.12) and (3.13) can be regarded as the generalized or extended likelihood equation under mixture \( f(x; Q) \). When \( g(\phi) = 1 \), and the mixing distribution is a single point distribution, i.e., the mixture problem degenerates into one parameter problem, eq. (3.13) is reduced into the parametric likelihood equation

\[
\sum_{i=1}^{n} \frac{\partial f(x_i; \phi)}{\partial \phi} = 0.
\]

By Theorem 3.4, \( \hat{Q} \) has the form

\[
\hat{Q}(\phi) = \sum_{j=1}^{D} \pi_j \{ \phi_j \}(\phi), \quad \pi_j \geq 0, \quad \sum_{j=1}^{D} \pi_j = 1, \quad \phi_1 < \phi_2 < \cdots < \phi_D.
\]

If \( 1 < j < D \), then let

\[
g(\phi) = \begin{cases} 
\frac{2(\phi - \phi_j + \phi_{j-1})}{\phi_j - \phi_{j-1}}, & \phi_j + \phi_{j-1} < \phi \leq \phi_j, \\
-\frac{2(\phi - \phi_{j-1})}{\phi_j + \phi_{j-1}} + 1, & \phi_j < \phi < \frac{\phi_{j+1} + \phi_j}{2}, \\
0, & \text{elsewhere.}
\end{cases}
\]

If \( j = 1 \), then let

\[
g(\phi) = \begin{cases} 
-\frac{2(\phi - \phi_1)}{\phi_2 - \phi_1} + 1, & \phi_1 \leq \phi < \frac{\phi_1 + \phi_2}{2}, \\
0, & \text{elsewhere.}
\end{cases}
\]

If \( j = D \), then let

\[
g(\phi) = \begin{cases} 
\frac{2(\phi - \phi_D + \phi_{D-1})}{\phi_D - \phi_{D-1}}, & \phi_D + \phi_{D-1} < \phi \leq \phi_D, \\
0, & \text{elsewhere.}
\end{cases}
\]

Then eq. (3.13) becomes

\[
\sum_{i=1}^{n} \pi_j \frac{\partial f(x_i; \phi_j)}{\partial \phi_j} = 0
\]

or

\[
\sum_{i=1}^{n} \frac{\pi_j \partial f(x_i; \phi_j)}{\sum_{j=1}^{D} \pi_j f(x_i; \phi_j)} = 0, \quad j = 1, \cdots, D. \tag{3.16}
\]
If \( j \neq D \) and we let
\[
g(\phi) = \begin{cases} 
\frac{1}{\pi_j}, & \phi = \phi_j, \\
-\frac{1}{\pi_D}, & \phi = \phi_D, \\
0, & \text{elsewhere,}
\end{cases}
\]
then eq.(3.12) becomes
\[
\sum_{i=1}^{n} \frac{f(x_i; \phi_j) - f(x_i; \phi_D)}{\sum_{j=1}^{D} f(x_i; \phi_j) \pi_j} = 0
\]
or
\[
\sum_{i=1}^{D} \frac{n_i f(x_i; \phi_j) - f(x_i; \phi_D)}{\sum_{j=1}^{D} \pi_j f(x_i; \phi_j)} = 0, \quad j = 1, \cdots, D - 1. \tag{3.17}
\]

Eq.(3.16) and (3.17) are simply the likelihood eq.(3.6) for the mixture.

### 3.6 Other Developments on MLEs for Mixing Distributions

#### 3.6.1 Consistency

General consistency results on the MLEs of mixing distributions are established by Kiefer and Wolfowitz (1956), Pfanzagl (1988) and Leroux (1992). Susko (1996) in his Ph.D. thesis compared the conditions given by them and showed that the conditions of Leroux (1992) imply the conditions of Kiefer and Wolfowitz (1956) as well as those of Pfanzagl (1988). The conditions of Pfanzagl are the weakest ones among them.

Taking into account Chen’s (1995) result that the optimal rate of convergence associated with the MLE for a finite mixing distribution with an unknown number of components is \( n^{-\frac{1}{2}} \), while the convergence rate \( n^{-\frac{1}{3}} \) is attainable when the number of components is known, improvement of the accuracy on the number of components of an estimator might speed convergence. Leroux (1992) proposed a maximum-penalized-likelihood method in which the estimator \( \hat{m} \) of the number of components maximizes
\[
l(L(\hat{Q}_m)) - a_m
\]
over $m$, where $a_m$ is a penalty term, satisfying $a_{m+1} \geq a_m$. This method discourages the selection of a model with an excessive number of components. Leroux (1992) proved that his maximum-penalized-likelihood estimator is also weakly consistent.

### 3.6.2 Binomial Mixtures

Based on Lindsay's (1983) work, Wood (1999) further elucidated the geometry underlying binomial mixtures. In the binomial mixture

$$p(x) = \int_0^1 f(x, \phi) dG(\phi), \quad x = 0, 1, \ldots, n,$$

the likelihood curve

$$\Gamma = \{L(\phi) = (f(0, \phi), f(1, \phi), \cdots, f(n, \phi)) : \phi \in \Omega\}$$

and the mixture set $\text{conv}(\Gamma)$ are all included in the simplex of multinomial distributions

$$P^n = \left\{p = (p_0, p_1, \cdots, p_n) : \sum_{i=0}^n p_i = 1, p_i \geq 0\right\},$$

a convex set with dimension $n$ in $R^{n+1}$. Here we assume all the results

$$x = 0, 1, \cdots, n$$

have been observed in the sample; this is always the case when the sample size $N$ is large enough.

We already know the maximizer of the log likelihood function

$$l = \sum_{i=0}^n n_i \log p_i$$

over $P^n$ is the empirical distribution

$$\hat{p} = \left(\frac{n_1}{N}, \cdots, \frac{n_n}{N}\right).$$
(i) If \( \hat{p} \) is not in the relative interior of \( \text{conv}(\Gamma) \), then the MLE for a mixing distribution is unique.

(ii) If \( \hat{p} \) is in the relative interior of \( \text{conv}(\Gamma) \), then the MLE for a mixing distribution is not unique.

We should note that if \( \hat{p} \) belongs to the relative interior of \( \text{conv}(\Gamma) \), Lindsay's (1983) result

\[
\hat{L} \in \partial \text{conv}(\Gamma)
\]

is not contradicted because for this special case,

\[
\text{conv}(\Gamma) = \partial \text{conv}(\Gamma)
\]

as an \( n \) dimensional subset of \( \mathbb{R}^{n+1} \).

Wood (1999) gave a simulation result that demonstrated that when \( n \) increases or \( N \) decreases, \( \hat{p} \) is less likely to lie in the relative interior of \( \text{conv}(\Gamma) \). He used \( \hat{p} \) to obtain the maximum likelihood vector \( \hat{L} \) on the mixture set by minimizing the Kullback-Leibler distance between \( \hat{p} \) and \( p \) in \( \text{conv}(\Gamma) \). Practical estimation begins with a discrete approximation to \( \Gamma \) by \( k \) piecewise linear segments. He showed that the resulting solution \( \hat{L}_k \) converges to \( \hat{L} \) in \( \mathbb{R}^{n+1} \) as \( k \to \infty \).

Obviously, Wood's (1999) method is heuristic. However, he did not theoretically resolve the computational problem in searching for the MLE for a mixing distribution, since minimizing Kullback-Leibler distance

\[
K(\hat{p}, L) = \sum_{i=0}^{n} \hat{p}_i \log \frac{\hat{p}_i}{L_i}
\]

between \( \hat{p} \) and \( L \) over \( \text{conv}(\Gamma) \) is essentially maximizing the log likelihood function on \( \text{conv}(\Gamma) \).
Chapter 4

Uniqueness of the NPMLE for a Mixing Distribution

4.1 Introduction

Consider a sequence of independent observations $x_1, x_2, \ldots, x_n$ from the mixture density

$$f(x; Q) = \int f(x; \theta) dQ(\theta),$$

where $f(x; \theta)$ are component densities on $\mathcal{X}$ and $Q$ is a probability measure on the parameter space $\Omega$.

In this chapter, we assume $\mathcal{X} \subset \mathbb{R}$ and $\Omega = [u, v] \subset \mathbb{R}$. Suppose the observation vector takes $D$ distinct values $y_1 < y_2 < \cdots < y_D$. Let $n_i$ be the number of observations which equal $y_i$, $\hat{p}(y_i) = \frac{n_i}{n}$, and $\hat{p}$ be the vector $(\frac{n_1}{n}, \ldots, \frac{n_D}{n})'$. The likelihood vector is

$$L(\theta) = (L_1(\theta), \ldots, L_D(\theta))$$

$$= (f(y_1; \theta), \ldots, f(y_D; \theta)),$$

thus the likelihood curve is

$$\Gamma = \{L(\theta) : \theta \in \Omega\}.$$
The likelihood mixture set is then the convex hull of \( \Gamma \), i.e., \( \text{conv}(\Gamma) \).

In Chapter 3, we have studied the geometric properties of the estimator \( \dot{Q} \) which maximizes the log likelihood

\[
l(Q) = \sum_{i=1}^{n} \ln f(x_i; Q) = \sum_{i=1}^{D} n_i \ln L_i(Q).
\]

Our objective in this section is to establish the uniqueness conditions for \( \dot{Q} \). Obviously, these conditions depend on the data. While the uniqueness conditions depend on the data, those for identifiability depend only on the model itself. This is the difference between the uniqueness and the identifiability. Clearly the uniqueness of the MLE \( \dot{Q} \) does not imply the identifiability of the model. Conversely, since the data is merely a part of the sample space, from the identifiability we also cannot conclude the uniqueness of \( \dot{Q} \).

There are a few authors considering the uniqueness and the support size of the NPMLE of a mixing distribution. Simar (1976) studied the uniqueness for mixtures of Poisson distributions and Jewell (1982) dealt with the uniqueness in exponential mixtures. Lindsay (1993) gave completely general results for the uniqueness of \( \dot{Q} \).

We already know that there exists a unique vector \( \hat{L} \in \partial\text{conv}(\Gamma) \) which maximizes the log likelihood \( l(L) \). Let \( Q \) be the set of mixing distributions \( Q \) which satisfy

\[
\hat{L} = L(Q) = \int_{\Omega} L(\theta) dQ(\theta).
\]

Of interest is the question of when the set \( Q \) has a single element. From Chapter 3 we know that all \( Q \) in \( Q \) must have support points in the set of zeroes of the Gateaux derivative \( l'(\hat{L}, L(\theta)) \), say \( Z = \{\hat{\theta}_1, \cdots, \hat{\theta}_s\} \). Lindsay (1993) made use of the variation diminishing property of the totally positive kernel to bound the cardinality of the set \( Z \).
In this chapter, we give the limits of the range of $Z$, and we prove that under the assumption of strictly total positivity and discreteness of $f(x; \theta)$, if $l'(\hat{L}, L(\theta)) = 0$ for all $\theta \in \Omega$, then $f(x; \theta)$ must be supported on $\{y_1, \ldots, y_D\}$, and it must be the case that $\hat{L} = \hat{p}$. This result ensures the uniqueness of the mixtures of a large number of discrete families of distributions. The uniqueness theorem of Simar (1976) for Poisson mixtures is a consequence of our result.

4.2 Totally Positive Kernels

Definition 4.1. A real function (frequently called kernel) $K(x, y)$ of two variables ranging over linearly ordered sets $X$ and $Y$, respectively, is said to be totally positive of order $r$ (abbreviated TP$_r$) if for all

$$x_1 < \cdots < x_m, \quad y_1 < \cdots < y_m, \quad x_i \in X, \quad y_j \in Y; \quad 1 \leq m \leq r,$$

we have the inequalities

\[
\begin{vmatrix}
K(x_1, y_1) & K(x_1, y_2) & \cdots & K(x_1, y_m) \\
K(x_2, y_1) & K(x_2, y_2) & \cdots & K(x_2, y_m) \\
\vdots & \vdots & \ddots & \vdots \\
K(x_m, y_1) & K(x_m, y_2) & \cdots & K(x_m, y_m)
\end{vmatrix} \geq 0.
\]

If strict inequality holds, then we say that $K(x, y)$ is strictly totally positive of order $r$ (STP$_r$) (See Karlin, 1968)

By Definition 4.1, TP$_r$ kernels are nonnegative functions and STP$_r$ kernels are positive functions. In what follows, if the subscript is omitted in the definition, then the property in question will be understood to hold for all values of $r$.

The class of totally positive functions includes most of the practically important density families in statistical theory. For example, it includes the exponential family, the noncentral $\chi^2$-distributions and the noncentral t-distributions. Another
thing we should note is that if $K(x,y)$ is $STP$, then for any $x_1 < \cdots < x_n$, $x_i \in X$, $K(x_1,y), \cdots, K(x_n,y)$ constitute a Tchebycheff system.

An important feature of totally positive kernels is their variation diminishing property (abbreviated $VDP$). Let $f(t)$ be defined in $I$, where $I$ is an order set of the real line. Let

$$S^-(f) = \sup S^-[f(t_1), \cdots, f(t_m)],$$

where the supremum is extended over all sets $t_1 < \cdots < t_m$, $t_i \in I$, $m$ is arbitrary but finite, and $S^-(x_1, \cdots, x_m)$ is the number of sign changes of the indicated sequence, with zero terms being discarded. A more refined method of counting sign changes is given by

$$S^+(f) = \sup S^+[f(t_1), \cdots, f(t_m)],$$

where $S^+(x_1, \cdots, x_m)$ denotes the maximum number of sign changes of the indicated sequence, the zero terms being permitted to take on arbitrary signs.

Let $K(x,y)$ defined on $X \times Y$ be Borel measurable, and assume for simplicity that the integral

$$\int_Y K(x,y)d\mu(y)$$

exists for every $x$ in $X$. Here $\mu$ represents a $\sigma$-finite measure on $R$ such that $\mu(U) > 0$ for each open set $U$ for which $U \cap Y$ is nonempty. Let $f$ be bounded and Borel measurable on $Y$, and consider the transformation

$$g(x) = (Tf)(x) = \int_Y K(x,y)f(y)d\mu(y).$$

**Lemma 4.1.** ($VDP$) (a) If $K$ is $TP_r$ and satisfies the integrability requirements stated above, then

$$S^-(g) = S^-(Tf) \leq S^-(f),$$

provided $S^-(f) \leq r - 1$. 
In the case in which \( f \) is piecewise continuous and \( S^-(f) = S^-(g) \leq r - 1 \), we further assert that the values of the functions \( f \) and \( g \) exhibit the same sequence of signs when their respective arguments traverse the domain of definition from left to right.

(b) Under the stronger condition that \( K \) is STP, and \( f \neq 0 \text{ a.e. } \mu \), we have

\[
S^+(g) = S^+(Tf) \leq S^-(f), \text{ provided } S^-(f) \leq r - 1.
\]

### 4.3 Preliminary Lemmas

In order to establish the uniqueness of the NPMLE for a mixing distribution, we make two assumptions:

**Assumption A.** Let \( \mathcal{X} \) be the support of the density or mass function \( f(x; \theta) \). It is assumed that if either \( \inf \mathcal{X} \) or \( \sup \mathcal{X} \) is finite and among the observed values, then a parametrization is available such that \( f(x; \theta) \) is defined and differentiable in \( \theta \) at those values of \( x \), with a unique maximum in \( \theta \).

**Assumption B.** The parameter space \([u, v]\) satisfies

\[
\tilde{\theta}_i \in [u, v], \quad i = 1, \ldots, D,
\]

where \( \tilde{\theta}_i \) maximizes \( f(y_i; \theta) \) over \( \theta \).

The following lemmas provide the necessary techniques to prove the uniqueness.

**Lemma 4.2.** If \( f(x; \theta) \) is STP, then

(a) For any fixed \( x_0 \in \mathcal{X} \), \( f(x_0; \theta) \) has a unique maximum point \( \tilde{\theta}_0 \).

(b) If \( \inf \mathcal{X} < x_0 < \sup \mathcal{X} \), then the derivative of \( f(x_0; \theta) \) at \( \tilde{\theta}_0 \) is zero;

If \( x_0 = \inf \mathcal{X} \), then \( \tilde{\theta}_0 = u \) and the right derivative of \( f(x_0; \theta) \) at \( \tilde{\theta}_0 = u \) is less than
or equal to zero; while if \( x_0 = \sup \mathcal{X} \), then \( \tilde{\theta}_0 = v \) and the left derivative of \( f(x_0; \theta) \) at \( \tilde{\theta}_0 = v \) is greater than or equal to zero.

Note that for this result, Lindsay’s (1993) proof is not accurate.

**Proof.** For any \( c > 0 \),

\[
f(x_0; \theta) - c = f(x_0; \theta) - c \int f(x; \theta) d\mu(x) = \int a(x)f(x; \theta)d(\mu + \delta_{x_0})(x),
\]

where the probability measure \( \delta_{x_0} \) has mass one at \( x_0 \) and

\[
a(x) = \begin{cases} 
-\frac{c\mu(\{x_0\})+1}{1+\mu(\{x_0\})}, & x = x_0, \\
-c, & x \neq x_0. 
\end{cases}
\]

(4.1)

The function \( a(x) \) has at most two sign changes, and the signs are in the order \((-,-,+)\). From Lemma 4.1,

\[
S^- (f(x_0; \theta) - c) \leq 2
\]

and the signs of \( f(x_0; \theta) - c \) are in the order \((-,-,+)\) when two sign changes occur. So \( f(x_0; \theta) \) is a unimodal function in \( \theta \). From Lemma 4.1, we also have

\[
S^+(f(x_0; \theta) - c) \leq 2,
\]

therefore \( f(x_0; \theta) \) has a unique maximum point \( \tilde{\theta}_0 \) and its derivative at \( \tilde{\theta}_0 \) is zero unless \( x_0 = \inf \mathcal{X} \) or \( x_0 = \sup \mathcal{X} \).

If \( x_0 = \inf \mathcal{X} \), from (4.1), \( a(x) \) has at most one sign change in \( x \), and the signs are in the order \((+,+)\) when it occurs. By Lemma 4.1, \( f(x_0; \theta) - c \) has at most one sign change in \( \theta \), and the sign order is \((+,+)\) when it occurs. So we conclude \( \tilde{\theta}_0 = u \) and the right derivative of \( f(x_0; \theta) \) at \( \tilde{\theta}_0 = u \) satisfies

\[
\frac{d}{d\theta} f(x_0; \theta)|_{\theta = \tilde{\theta}_0 = u} \leq 0.
\]
The same argument applies to $x_0 = \sup X$; specifically we infer that $\bar{\theta}_0 = v$ and the left derivative of $f(x_0; \theta)$ at $\bar{\theta}_0 = v$ satisfies
\[
\frac{d}{d\theta} f(x_0; \theta)\big|_{\theta = \bar{\theta}_0 = v} \geq 0. \quad \square
\]

In order to state the results that follow below, we need to define the function $S(y_1, \cdots, y_D)$ for the data sequence $y_1 < \cdots < y_D$, $y_i \in X$:
\[
S(y_1, \cdots, y_D) = I_{\{y_1 > \inf X\}} + I_{\{y_D < \sup X\}} + 2 \sum_{i=1}^{D-1} I_{\{(y_i, y_{i+1}) \cap X \neq \emptyset\}},
\]
where
\[
I_A = \begin{cases} 1, & A \text{ occurs,} \\ 0, & \text{otherwise.} \end{cases}
\]

**Lemma 4.3.** If $f(x; \theta)$ is STP, then
\[
S^+(l'(\hat{L}, L(\theta))) \leq S(y_1, \cdots, y_D)
\]
unless $l'(\hat{L}, L(\theta)) = 0$ for all $\theta$.

**Proof.** If $L(\theta^0)$ is on the support hyperplane $H$ of $\text{conv}(\Gamma)$ at $\hat{L}$, then
\[
l'(\hat{L}, L(\theta^0)) = \sum_{i=1}^{D} n_i \left( \frac{L_i(\theta^0)}{\hat{L}_i} - 1 \right)
= \sum_{i=1}^{D} n_i \left( \frac{f(y_i; \theta^0)}{\hat{L}_i} - 1 \right)
= 0.
\]
So
\[
\sum_{i=1}^{D} \frac{\hat{p}(y_i) f(y_i; \theta^0)}{\hat{L}_i} = 1.
\]
Denote $F_1$ as the discrete probability distribution function with mass
\[
dF_1(y_i) = \frac{\hat{p}(y_i) f(y_i; \theta^0)}{\hat{L}_i}
\]
at $y_i$.

If $F_2$ is the probability distribution function corresponding to density $f(x; \theta^0)$, then

$$l'(\tilde{L}, L(\theta)) = n \int_X \frac{f(x; \theta)}{f(x; \theta^0)} d\{F_1(x) - F_2(x)\}.$$ 

Define the $\sigma$–finite measure $\nu$ by

$$d\nu(x) := d\{F_1(x) + F_2(x)\}.$$ 

In addition, define the functions

$$f_1(x) := \begin{cases} \frac{dF_1(x)}{d(F_1(x) + F_2(x))}, & \text{if } x \in \{y_1, \cdots, y_D\}, \\ 0, & \text{otherwise}, \end{cases}$$

$$f_2(x) := \begin{cases} \frac{dF_2(x)}{d(F_1(x) + F_2(x))}, & \text{if } x \in \{y_1, \cdots, y_D\}, \\ 1, & \text{otherwise}. \end{cases}$$

(4.2) (4.3)

$f_1$ and $f_2$ are the Radon-Nikodym derivatives of $F_1$ and $F_2$ with respect to $\nu$. Hence

$$l'(\tilde{L}, L(\theta)) = n \int_X \frac{f(x; \theta)}{f(x; \theta^0)} [f_1(x) - f_2(x)] d\nu(x).$$

(4.4)

In eq.(4.4), the kernel

$$n \frac{f(x; \theta)}{f(x; \theta^0)}$$

is $STP$, and

$$f_1(x) - f_2(x) = -1$$

on $X \setminus \{y_1, \cdots, y_D\}$. At the observations $y_i$, $i = 1, \cdots, D$, this difference can be positive, so

$$f_1(x) - f_2(x)$$

has at most $S(y_1, \cdots, y_D)$ sign changes, that is,

$$S^-(f_1 - f_2) \leq S(y_1, \cdots, y_D).$$
Applying Lemma 4.1(b) to the transformation \( l'(\hat{\mathbf{L}}, \mathbf{L}(\theta)) \), if

\[ l'(\hat{\mathbf{L}}, \mathbf{L}(\theta)) \neq 0, \]

then

\[ f_1 - f_2 \neq 0, \text{ a.e.v.} \]

So

\[ S^+(l'(\hat{\mathbf{L}}, \mathbf{L}(\theta))) \leq S^-(f_1 - f_2) \leq S(y_1, \ldots, y_D). \]

\[ \square \]

Lemma 4.4. If \( f(x; \theta) \) is STP, then for \( \tilde{\theta}_i \) which maximizes \( f(y_i; \theta) \) over \( \theta, i = 1, \ldots, D, \)

\[ \tilde{\theta}_1 < \tilde{\theta}_2 < \cdots < \tilde{\theta}_D. \]

Proof. By Lemma 4.2

\[ f(y_1; \tilde{\theta}_1) > f(y_1; \tilde{\theta}_2), \]

\[ f(y_2; \tilde{\theta}_2) > f(y_2; \tilde{\theta}_1), \]

so

\[ \begin{vmatrix} f(y_1; \tilde{\theta}_1) & f(y_2; \tilde{\theta}_1) \\ f(y_1; \tilde{\theta}_2) & f(y_2; \tilde{\theta}_2) \end{vmatrix} > 0. \]

Then \( y_1 < y_2 \) and the STP property of \( f(x; \theta) \) imply \( \tilde{\theta}_1 < \tilde{\theta}_2 \). Consequently, we have

\[ \tilde{\theta}_1 < \tilde{\theta}_2 < \cdots < \tilde{\theta}_D. \]

\[ \square \]

The following lemma gives the limits of the range of \( Z \), the set of zeros of \( l'(\hat{\mathbf{L}}, \mathbf{L}(\theta)) \).

Lemma 4.5. If \( f(x; \theta) \) is STP, then

(a) For all \( \tilde{\theta}_i \in Z \), \( \tilde{\theta}_1 \leq \tilde{\theta}_i \leq \tilde{\theta}_D. \)

(b) If \( l'(\hat{\mathbf{L}}, \mathbf{L}(\tilde{\theta}_1)) = 0 \), then \( \tilde{\theta}_1 = u, y_1 = \inf \mathcal{X} \). If \( l'(\hat{\mathbf{L}}, \mathbf{L}(\tilde{\theta}_D)) = 0 \), then \( \tilde{\theta}_D = v, y_D = \sup \mathcal{X} \).
Proof. (a) We will show that \( l'(\tilde{\mathbf{L}}, \mathbf{L}(\theta_0)) = 0 \) implies \( \theta_0 \geq \tilde{\theta}_1 \). If \( \theta_0 < \tilde{\theta}_1 \), then

\[
\frac{d}{d\theta^+} l'(\tilde{\mathbf{L}}, \mathbf{L}(\theta))|_{\theta=\theta_0} > 0.
\]

Note that

\[
\frac{d}{d\theta^+} l'(\tilde{\mathbf{L}}, \mathbf{L}(\theta))|_{\theta=\theta_0} = \sum_{i=1}^{D} \frac{n_i}{L_i} \frac{d}{d\theta^+} L_i(\theta)|_{\theta=\theta_0}
\]

\[
= \sum_{i=1}^{D} \frac{n_i}{L_i} \frac{d}{d\theta^+} f(y_i; \theta)|_{\theta=\theta_0}.
\]

In this expression,

\[
\frac{n_i}{L_i} > 0,
\]

and by the proof of Lemma 4.2, \( f(y_i; \theta) \) is unimodal and

\[
S^+(f(y_i; \theta) - c) \leq 2
\]

for any \( c > 0 \). We also know \( \theta_0 \) is on the left side of the maximum point \( \tilde{\theta}_i \); therefore

\[
\frac{d}{d\theta^+} f(y_i; \theta)|_{\theta=\theta_0} > 0.
\]

Since

\[
l'(\tilde{\mathbf{L}}, \mathbf{L}(\theta_0)) = 0,
\]

\[
l'(\tilde{\mathbf{L}}, \mathbf{L}(\theta)) > 0
\]

for all \( \theta \) in the right neighborhood of \( \theta_0 \), which contradicts the characterization of \( \tilde{\mathbf{L}} \). Therefore \( \theta_0 \geq \tilde{\theta}_1 \). Similar arguments show that

\[
l'(\tilde{\mathbf{L}}, \mathbf{L}(\theta_0)) = 0
\]

implies \( \theta_0 \leq \tilde{\theta}_D \).

(b) If \( l'(\tilde{\mathbf{L}}, \mathbf{L}(\tilde{\theta}_1)) = 0 \), then by the characterization of \( \tilde{\mathbf{L}} \), we must have

\[
\frac{d}{d\theta^+} l'(\tilde{\mathbf{L}}, \mathbf{L}(\theta))|_{\theta=\tilde{\theta}_1} \leq 0.
\]
Note that 
\[ \frac{d}{d\theta^+} l'([\hat{L}, L(\theta)])|_{\theta=\hat{\theta}_1} = \sum_{i=1}^{D} \frac{n_i}{L_i} \frac{d}{d\theta^+} f(y_i; \theta)|_{\theta=\hat{\theta}_1} \]
and
\[ \frac{n_i}{L_i} > 0, \quad \frac{d}{d\theta^+} f(y_i; \theta)|_{\theta=\hat{\theta}_1} > 0, \quad i = 2, \ldots, D, \]
\[ \frac{d}{d\theta^-} f(y_1; \theta)|_{\theta=\hat{\theta}_1} < 0. \]
Therefore \( \hat{\theta}_1 = u \) and by Lemma 4.2, \( y_1 = \inf X \). The assertion about \( \hat{\theta}_D \) can be proved in the similar way.

\[ \square \]

4.4 Uniqueness of the NPMLE for a Mixing Distribution

Let
\[ c = I_{\{y_1=\inf X\}} + I_{\{y_D=\sup X\}}. \]

The following theorem is from Lindsay (1993).

**Theorem 4.1.** Suppose \( f(x; \theta) \) is STP. If
\[ l'(\hat{L}, L(\theta)) \neq 0, \]
then the MLE \( \hat{Q} \) is unique, with no more than \( \frac{1}{2} \{ S(y_1, \ldots, y_D) + c \} \) points of support.

Theorem 4.1 asserts that in the STP class, if the Gateaux derivative
\[ l'(\hat{L}, L(\theta)) \neq 0, \]
then the MLE \( \hat{Q} \) is unique. So we know that uniqueness may fail only when
\[ l'(\hat{L}, L(\theta)) \equiv 0. \]
We further conclude that when

\[ l'(\hat{L}, L(\theta)) = 0, \]

the density \( f(x; \theta) \) must be supported on \( \{y_1, y_2, \ldots, y_D\} \).

**Theorem 4.2.** If \( f(x; \theta) \) is STP, then

(a) \[ l'(\hat{L}, L(\theta)) = 0 \]

for all \( \theta \) if and only if

\[ f(y; Q^0) = \hat{p}(y) \]

for all \( Q^0 \in Q \), i.e., for \( Q^0 \) such that

\[ L(Q^0) = \hat{L}. \]

(b) If

\[ l'(\hat{L}, L(\theta)) = 0 \]

for all \( \theta \) and \( f(x; \theta) \) is a discrete distribution, then \( f(x; \theta) \) is supported on \( \{y_1, y_2, \ldots, y_D\} \) for all \( \theta \).

**Proof.** (a) From (4.2), (4.3) and (4.4) in the proof of Lemma 4.3 and the VDP in Lemma 4.1, we may conclude

\[ l'(\hat{L}, L(\theta)) = 0 \]

if and only if

\[ f_1(x) - f_2(x) = 0. \]

The later holds if and only if the measure

\[ \nu(\mathcal{X} \setminus \{y_1, \ldots, y_D\}) = 0 \]
and
\[ \hat{L}_i = L_i(Q^0) = f(y_i; Q^0) = \hat{p}(y_i) \]
for all \( Q^0 \in Q \), which completes the proof of Theorem 4.2 (a).

(b) If
\[ l'(\hat{L}, L(\theta)) = 0 \]
for all \( \theta \), then from (a) we know
\[ \hat{L} = \hat{p}. \]
Thus the gradient of \( l(p) \) at \( \hat{L} \) is the \( D \)-dimensional vector
\[ \nabla l(\hat{L}) = \nabla l(\hat{p}) = (n, n, \cdots, n), \]
which is orthogonal to the probability simplex!

Hence the plane in which the probability simplex resides is exactly the support hyperplane of \( \text{conv}(\Gamma) \) at \( \hat{L} \) (or \( \hat{p} \)), and the whole likelihood curve \( \Gamma \) resides in this plane. So \( f(x; \theta) \) is a discrete distribution with support contained in \( \{y_1, \cdots, y_D\} \).

From Theorem 4.2(b) and Theorem 4.1, we learn all mass families in the \( STP \) class with support not in \( \{y_1, \cdots, y_D\} \) must have a unique mixture MLE. This leads to the following corollary.

**Corollary 4.1.** For any discrete probability mass family \( f(x; \theta) \) which is \( STP \) and whose support is not included in \( \{y_1, \cdots, y_D\} \) for some \( \theta \), the mixture NPMLE \( \hat{Q} \) is unique, with no more than \( \frac{1}{2} \{S(y_1, \cdots, y_D) + c\} \) support points.

We have already learned that the \( STP \) class is very extensive; it includes many probability mass families that we are interested in, such as the Poisson family
\[ f(n; \lambda) = \frac{\lambda^n}{n!} e^{-\lambda}, \quad n = 0, 1, \cdots, \lambda > 0. \]
We know when \( x_1 < x_2 < \cdots < x_n \) the matrix
\[
\begin{pmatrix}
1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\
1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & x_n & x_n^2 & \cdots & x_n^{n-1}
\end{pmatrix}
\]
is positive definite for any \( n \geq 1 \). So for
\[
n_1 < n_2 < \cdots < n_m, \quad \lambda_1 < \lambda_2 < \cdots < \lambda_m,
\]
\[
\begin{vmatrix}
\lambda_1^{n_1} & \lambda_1^{n_2} & \cdots & \lambda_1^{n_m} \\
\lambda_2^{n_1} & \lambda_2^{n_2} & \cdots & \lambda_2^{n_m} \\
\vdots & \vdots & \ddots & \vdots \\
\lambda_m^{n_1} & \lambda_m^{n_2} & \cdots & \lambda_m^{n_m}
\end{vmatrix} > 0.
\]
Hence,
\[
\begin{vmatrix}
f(n_1; \lambda_1) & f(n_2; \lambda_1) & \cdots & f(n_m; \lambda_1) \\
f(n_1; \lambda_2) & f(n_2; \lambda_2) & \cdots & f(n_m; \lambda_2) \\
\vdots & \vdots & \ddots & \vdots \\
f(n_1; \lambda_m) & f(n_2; \lambda_m) & \cdots & f(n_m; \lambda_m)
\end{vmatrix} = \begin{vmatrix}
\lambda_1^{n_1}e^{-\lambda_1} & \lambda_2^{n_1}e^{-\lambda_1} & \cdots & \lambda_m^{n_1}e^{-\lambda_1} \\
\lambda_1^{n_2}e^{-\lambda_2} & \lambda_2^{n_2}e^{-\lambda_2} & \cdots & \lambda_m^{n_2}e^{-\lambda_2} \\
\vdots & \vdots & \ddots & \vdots \\
\lambda_1^{n_m}e^{-\lambda_m} & \lambda_2^{n_m}e^{-\lambda_m} & \cdots & \lambda_m^{n_m}e^{-\lambda_m}
\end{vmatrix}
\]
\[
= e^{-(\lambda_1 + \cdots + \lambda_m)} \begin{vmatrix}
\lambda_1^{n_1} & \lambda_2^{n_1} & \cdots & \lambda_m^{n_1} \\
\lambda_2^{n_2} & \lambda_2^{n_2} & \cdots & \lambda_m^{n_2} \\
\vdots & \vdots & \ddots & \vdots \\
\lambda_m^{n_m} & \lambda_2^{n_m} & \cdots & \lambda_m^{n_m}
\end{vmatrix} > 0,
\]
that is, \( f(n; \lambda) \) is STP. Combining Corollary 4.1, a Poisson mixture has a unique MLE, which coincides with the uniqueness theorem in Simar (1976).

From Theorem 4.2(a), if
\[
l'(\hat{\mathbf{L}}, \mathbf{L}(\theta)) \equiv 0,
\]
then the model fits the data perfectly. By the discussion in Chapter 2 on the identifiability of mixture models, in such a case, the uniqueness occurs if and only if \( \hat{\mathbf{p}} \) is on the relative boundary of \( \text{conv}(\Gamma) \).
Corollary 4.2. Let $f(x; \theta)$ be STP. If

$$l'(\hat{L}, L(\theta)) = 0$$

for all $\theta$, then the MLE $\hat{Q}$ of the mixing distribution is unique if and only if

(a) $\hat{p}$ is on the relative boundary of $\text{conv}(\Gamma)$; or

(b) $\text{index}(\hat{p}) \leq \frac{D-1}{2}$. 

Chapter 5

Method of Moments Estimation of Mixing Distributions

5.1 Introduction

If a probability distribution function $F(x)$ possesses finite moments of all orders, then we can calculate its moments $m_n$ by the integrals

$$m_n = \int_{-\infty}^{\infty} x^n dF(x), \quad n = 0, 1, \ldots.$$ 

Conversely, we can ask under what kind of conditions a probability distribution function $F(x)$ is completely determined by its moments $(m_n)_{n=0}^\infty$?

This question is closely related to the Tchebycheff inequality. Some sufficient conditions, necessary and sufficient conditions can be found in Ahiezer (1965). A frequently used criterion can be found in Cramér (1946), which states that a probability distribution $F$ is uniquely determined by its moments $m_n$ whenever the power series

$$\sum_{n=0}^{\infty} \frac{m_n t^n}{n!}$$

is absolutely convergent for some $t > 0$.

Consequently every probability distribution with support contained in $[0, 1]$ is totally determined by its moment sequence (See Shohat and Tamarkin (1963)). Feller
(1971) recovered a probability distribution function $F(x)$ supported on $[0,1]$ from its moment sequence $(m_n)_{n=0}^\infty$ and gave an explicit formula for $F$:

$$F(x) = \lim_{n \to \infty} \sum_{i \leq nx} \binom{n}{i} (-1)^{n-i} \Delta^{n-i} m_i$$

for every continuous point $x$ of $F$, where the differences $\Delta^r m_i$ are defined by

$$\Delta^r m_i = \sum_{j=0}^r \binom{r}{j} (-1)^{r-j} m_{i+j}, \quad r = 1, 2, \ldots$$

and $\Delta^0 m_i = m_i$.

In section 5.2.1, we study the $(n+1)$-point distribution with one given support point that fits the first $2n+1$ prescribed moments $m_0, m_1, \ldots, m_{2n}$. We then utilize the Tchebycheff inequality given in Ahiezer (1965) and Lindsay and Basak (2000) to reveal the relationship among the determination of the moment problem, the continuity (or discontinuity) of the underlying distribution function and the limit of the window function.

Suppose we want to estimate an unknown distribution function $F(x)$ that is uniquely determined by its moment sequence $(m_n)_{n=0}^\infty$. If for each $n$, $m_n$ has an estimate $\hat{m}_n$ such that $\hat{m}_n \to m_n$ almost surely as $N \to \infty$, then it is possible for us to make use of $(\hat{m}_n)_{n=0}^\infty$ to estimate $F(x)$.

The first problem we encounter is that $(\hat{m}_n)_{n=0}^\infty$ may not be a moment sequence of any distribution. In this case, we have to properly modify $(\hat{m}_n)_{n=0}^\infty$ into a moment sequence $(\hat{m}_n)_{n=0}^\infty$ such that the underlying probability distribution function $\hat{F}_N$ is supported on the required set and converges almost surely to $F$ weakly.

We consider a mixture model

$$\int_{\lambda \in \Lambda} f(x; \lambda) dG(\lambda),$$

where $f(x; \lambda)$ is a probability density (or mass) function with parameter $\lambda$ in $\Lambda$, and $G(\lambda)$ is a probability distribution function supported on $\Lambda$. 


In order to estimate $G$ along the lines described above, we often use the moments of a function of $\lambda$ chosen so that the moment sequence can be easily and consistently estimated. We present the estimation of normal mixtures as an example. Sometimes we first need to estimate a weighted distribution $H$ defined by $dH(\lambda) = w(\lambda)dG(\lambda)$, where $w(\lambda)$ is strictly positive, and then perform an appropriate transformation. We present the estimation of mixing distributions in negative binomial mixtures as an example.

Tucker (1963) initially introduced a moment method to estimate the mixing distribution in a Poisson mixture. He first estimated the weighted distribution

\[ dH(\lambda) = \frac{e^{-\lambda}}{p(0)} dG(\lambda), \]

and then performed a transform to estimate $G$. His estimator is weakly consistent. However, since $p(0)$ is used to replace $p(0)$, his estimator may not be a probability distribution, so a modification is needed. Tucker’s (1963) paper is a pioneer paper which throws light on the subject of estimating mixing distributions.

Brockett (1977) followed Tucker’s moment method obtaining a consistent nonparametric estimator of the mixing distributions in an exponential mixture and in a scale mixture of symmetric stable distributions.

Section 5.3 is devoted to the estimation of mixing distributions by the moment method. Specifically, the identifiability of a negative binomial mixture and a normal mixture is discussed. A consistent nonparametric estimate of the mixing distribution in a negative binomial mixture, a normal mixture and a binomial mixture is constructed along the line in Tucker (1963). Appealing to a result from Lindsay (1989), we give a simpler proof of weak convergence than that of Tucker (1963).
5.2 The Problem of Moments

The theory of the problem of moments is of fundamental importance. In addition, it also provides a powerful tool for estimating mixing distributions. In this section we derive some results concerning Tchebycheff inequalities. We also introduce several lemmas as preliminaries of construction of consistent estimates of mixing distributions.

5.2.1 Determination of the Problem of Moments, Continuity (or Discontinuity) of the Underlying Distribution Function and the Limit of Window Function

Generally we call a bounded nondecreasing function on \( \mathbb{R} \) a distribution function. A point \( x_0 \in \mathbb{R} \) is called a support point of the distribution function \( \varphi \), if for every open interval \( I \) containing \( x_0 \),

\[
\int_I d\varphi(x) > 0.
\]

The set of all support points of \( \varphi \) is called the support of \( \varphi \).

A sequence of real numbers \( m = (m_0, m_1, \ldots) \) is called a moment sequence if there exists a distribution function \( \varphi(x) \) on \( \mathbb{R} \) such that

\[
m_n = \int_{-\infty}^{\infty} x^n d\varphi(x), \quad n = 0, 1, \ldots, \quad (5.1)
\]

When the support of \( \varphi \) in (5.1) is contained in \((-\infty, \infty), (0, \infty) \) and \([0,1]\), \( m \) is called a Hamburger, Stieltjes and Hausdorff moment sequence respectively.

Given \( m_0, m_1, \ldots, m_n \) in (5.1), provided \( \varphi^{(n)} \) is any distribution function such that

\[
m_i = \int_{-\infty}^{\infty} x^i d\varphi^{(n)}(x), \quad i = 0, 1, \ldots, n,
\]

then what about

\[
\lim_{n \to \infty} |\varphi^{(n)}(x) - \varphi(x)|
\]
for each $x \in \mathbb{R}$? If this limit is 0 for any $x \in \mathbb{R}$, then the distribution $\varphi$ is completely determined by the moments $\mathbf{m} = (m_0, m_1, \cdots)$. In order to reply the above question, the Tchebycheff inequality was introduced. This inequality was simultaneously proved by Markoff and Stieltjes.

Let

$$\Delta_n(\mathbf{m}) := \begin{vmatrix} m_0 & m_1 & \cdots & m_n \\ m_1 & m_2 & \cdots & m_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ m_n & m_{n+1} & \cdots & m_{2n} \end{vmatrix} = \left|m_{i+j}\right|_{i,j=0}^n, \quad n = 0, 1, \ldots.$$ 

Suppose the first $2n + 1$ moments $m_0, m_1, \cdots, m_{2n}$ in (5.1) are given. Provided that

$$\Delta_0(\mathbf{m}) > 0, \quad \cdots, \quad \Delta_n(\mathbf{m}) > 0,$$

there exists the right continuous distribution function $\psi$ with these first $2n + 1$ moments which has exactly $n + 1$ support points and among them a given real number $\nu$ such that

$$P_n(\nu; \mathbf{m}) \neq 0,$$

with

$$P_n(\nu; \mathbf{m}) = \frac{c}{\Delta_{n-1}(\mathbf{m})} \begin{vmatrix} 1 & \nu & \cdots & \nu^n \\ m_0 & m_1 & \cdots & m_n \\ m_1 & m_2 & \cdots & m_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n-1} & m_n & \cdots & m_{2n-1} \end{vmatrix},$$

and where $c \neq 0$ is a constant. The polynomials

$$P_{-1}(x; \mathbf{m}), \quad P_0(x; \mathbf{m}), \quad P_1(x; \mathbf{m}), \quad \cdots$$

have relationship

$$P_{-1}(x; \mathbf{m}) = 0, \quad P_0(x; \mathbf{m}) = 1,$$

$$P_i(x; \mathbf{m}) = (\alpha_i x + \beta_i)P_{i-1}(x; \mathbf{m}) - P_{i-2}(x; \mathbf{m}), \quad i = 1, 2, \ldots \tag{5.2}$$
Given $m_0, m_1, \cdots, m_{2n}$, then

$$\{P_i(x; m)\}^{n}_{i=0}, \{\alpha_i\}^{n+1}_{i=1} \text{ and } \{\beta_i\}^{n}_{i=1}$$

can be determined (See Uspensky (1937, page 366)). The $n + 1$ support points $\xi_1 < \xi_2 < \cdots < \xi_{n+1}$ of $\psi$ are the roots of the polynomial

$$P(x; m) = \left(\alpha_{n+1}(x - \nu) - \frac{P_{n-1}(\nu; m)}{P_n(\nu; m)}\right)P_n(x; m) - P_{n-1}(x; m).$$

If $\nu$ is a root of $P_{n+1}(x; m)$, then

$$P(x; m) = P_{n+1}(x; m). \tag{5.3}$$

Taking into account (5.2) and

$$0 = P_{n+1}(\nu; m) = P(\nu; m) = \left(\alpha_{n+1}\nu - \frac{P_{n-1}(\nu; m)}{P_n(\nu; m)}\right)P_n(\nu; m) - P_{n-1}(\nu; m),$$

by the interwoven property of roots to $P_n(x; m)$ and $P_{n+1}(x; m)$ (Lindsay (1989), Lemma 2D),

$$P_n(\nu; m) \neq 0,$$

hence

$$\beta_{n+1} = -\alpha_{n+1}\nu - \frac{P_{n-1}(\nu; m)}{P_n(\nu; m)},$$

so equality (5.3) is valid.

In Uspensky (1937), it is shown that

$$\psi(\xi_i + 0) \geq \varphi(\xi_i + 0) \geq \varphi(\xi_i - 0) \geq \psi(\xi_i - 0). \tag{5.4}$$

In fact we can further conclude that if $\varphi \neq \psi$, then

$$\psi(\xi_i + 0) > \varphi(\xi_i + 0) \geq \varphi(\xi_i - 0) > \psi(\xi_i - 0)$$

for $2 \leq i \leq n$. 
The Tchebycheff inequality states

\[ 0 \leq \varphi(\nu) - \psi(\nu - 0) \leq \psi(\nu + 0) - \psi(\nu - 0) = \frac{1}{P_{n+1}(\nu; m)P_n(\nu; m) - P'_n(\nu; m)P_{n+1}(\nu; m)}. \]

If \( \varphi_1(\nu) \) is another distribution function fitting the moments

\[ m_0, m_1, \ldots, m_{2n}, \]

then we also have

\[ 0 \leq \varphi_1(\nu) - \psi(\nu - 0) \leq \psi(\nu + 0) - \psi(\nu - 0) = \frac{1}{P_{n+1}(\nu; m)P_n(\nu; m) - P'_n(\nu; m)P_{n+1}(\nu; m)}. \]

hence

\[ |\varphi(\nu) - \varphi_1(\nu)| \leq \frac{1}{P_{n+1}(\nu; m)P_n(\nu; m) - P'_n(\nu; m)P_{n+1}(\nu; m)}. \]

Ahiezer(1965) stated the Tchebycheff inequality more clearly. He defined the window functions as

\[ w_n(x; m) = \left[ \sum_{k=0}^{n} |Q_k(x; m)|^2 \right]^{-1}, \]

where \( Q_k(x; m) \) are the orthogonal polynomials

\[ Q_0(x; m) = 1, \]
\[ Q_1(x; m) = \frac{m_1 - m_0x}{\sqrt{\Delta_1(m)}}, \]
\[ \ldots \]
\[ Q_k(x; m) = \frac{\hat{P}_k(x; m)}{\sqrt{\Delta_{k-1}(m)\Delta_k(m)}}, \quad k = 2, 3, \ldots, \]
with

\[
\hat{P}_k(x; m) = \begin{vmatrix}
1 & x & \cdots & x^k \\
m_0 & m_1 & \cdots & m_k \\
m_1 & m_2 & \cdots & m_{k+1} \\
\vdots & \vdots & \ddots & \vdots \\
m_{k-1} & m_k & \cdots & m_{2k-1}
\end{vmatrix}.
\]

Under this specification, \( \psi \) has mass \( w_n(\nu; m) \) at \( \nu \) and

\[
|\varphi(x) - \varphi_1(x)| \leq w_n(x; m), \quad x \in R. \tag{5.5}
\]

Furthermore, Lindsay (2000) proved that

\[
w_n(x; m) = \frac{1}{v_n'(x) M_n^{-1} v_n(x)},
\]

where \( M_n \) is the moment matrix

\[
M_n = \begin{pmatrix}
m_0 & m_1 & \cdots & m_n \\
m_1 & m_2 & \cdots & m_{n+1} \\
\vdots & \vdots & \ddots & \vdots \\
m_n & m_{n+1} & \cdots & m_{2n}
\end{pmatrix}
\]

and \( v_n(x) = (1, x, \cdots, x^n)' \). From (5.4) we know that if \( \varphi \) is continuous at \( \nu \), then

\[
[\varphi(\nu) - \psi(\nu - 0)] \lor [\psi(\nu) - \varphi(\nu)] \geq \frac{1}{2} w_n(\nu; m).
\]

This inequality together with the Tchebycheff inequality (5.5) leads to the following theorem.

**Theorem 5.1.** *If the solution \( \varphi \) of (5.1) is continuous, then \( \varphi \) is completely determined by \( m \) if and only if \( w_n(x; m) \to 0 \) for each \( x \in R \).*

We note that \( \{w_n(x; m)\}_{n=0}^{\infty} \) is a decreasing positive sequence, so it must converge to a nonnegative number. Thus we have the following

**Theorem 5.2.** *If \( \varphi \) is completely determined by \( m \), then

\[
\varphi(x_0) - \varphi(x_0 - 0) = c(x_0),
\]

where \( c(x_0) \) is a constant.*
where
\[ c(x_0) = \lim_{n \to \infty} w_n(x_0; m) \]

**Proof.** Since \( \varphi \) is determined by \( m \),
\[ \varphi(x_0 - 0) - \psi(x_0 - 0) \to 0 \]
and
\[ \psi(x_0 + 0) - \varphi(x_0 + 0) \to 0. \]
Hence,
\[ \varphi(x_0 + 0) - \varphi(x_0 - 0) = \varphi(x_0 + 0) - \psi(x_0 + 0) + \psi(x_0 + 0) - \psi(x_0 - 0) + \psi(x_0 - 0) - \varphi(x_0 - 0) \]
\[ \to c(x_0). \]
Thus,
\[ \varphi(x_0) - \varphi(x_0 - 0) = c(x_0). \]

According to Theorem 5.2, if \( \varphi \) is completely determined by \( m \), then
\[ c(x_0) > 0 \]
if and only if \( \varphi \) is discontinuous at \( x_0 \); and
\[ c(x_0) = 0 \]
if and only if \( \varphi \) is continuous at \( x_0 \).

### 5.2.2 Preliminary Lemmas

In this section, we introduce some lemmas that are useful for constructing consistent estimates of mixing distributions. For any sequence of real numbers \( m = (m_0, m_1, \cdots) \), we define \( \Delta_m^{(1)}(m) \) and \( \Delta_m^{(2)}(m) \) as
Lemma 5.1. Let $\mathbf{m} = (m_0, m_1, \cdots)$ be a sequence of real numbers.

(a) A necessary condition for $\mathbf{m}$ be a moment sequence is that

$$\Delta_n(\mathbf{m}) \geq 0, \quad n = 0, 1, \cdots.$$ 

(b) In order that there exists a distribution function $\varphi(x)$ which has infinite many support points such that

$$m_n = \int_{-\infty}^{\infty} x^n d\varphi(x), \quad n = 0, 1, \cdots,$$

it is necessary and sufficient that

$$\Delta_n(\mathbf{m}) > 0, \quad n = 0, 1, \cdots.$$ 

(c) In order that there exists a distribution function $\varphi(x)$ which has precisely $k + 1$ support points such that

$$m_n = \int_{-\infty}^{\infty} x^n d\varphi(x), \quad n = 0, 1, \cdots,$$

it is necessary and sufficient that

$$\Delta_i(\mathbf{m}) > 0, \quad 0 \leq i \leq k,$$

$$\Delta_i(\mathbf{m}) = 0, \quad i \geq k + 1.$$
In this case, \( \varphi(x) \) is uniquely determined by \( m \).

The following result is due to Mammana (1954).

**Lemma 5.2.** In order that for \( 2k+2 \) real numbers \( m = (m_0, m_1, m_2, \ldots, m_{2k+1}) \), there exist \( k + 1 \) distinct real numbers \( x_1, \ldots, x_{k+1} \) and \( k + 1 \) positive numbers \( p_1, \ldots, p_{k+1} \) such that

\[
\sum_{i=1}^{k+1} p_i x_i^n = m_n, \quad n = 0, 1, \ldots, 2k + 1,
\]

it is necessary and sufficient that

\[
\Delta_i(m) > 0 \quad \text{for} \quad 0 \leq i \leq k.
\]

This solution is necessarily unique. Furthermore, \( x_1, \ldots, x_{k+1} \) are the \( k + 1 \) distinct roots of the polynomial equation

\[
\begin{vmatrix}
1 & x & \cdots & x^{k+1} \\
m_0 & m_1 & \cdots & m_{k+1} \\
m_1 & m_2 & \cdots & m_{k+2} \\
\vdots & \vdots & \ddots & \vdots \\
m_k & m_{k+1} & \cdots & m_{2k+1}
\end{vmatrix} = 0,
\]

and \( p_1, \ldots, p_{k+1} \) are the solutions of the system of linear equations

\[
\sum_{i=1}^{k+1} p_i x_i^n = m_n, \quad n = 0, 1, \ldots, k.
\]

**Lemma 5.3.** (a) In order that there exists a solution \( \varphi \) of the moment problem (5.1) that has precisely \( n + 1 \) support points that are all in \( (0, \infty) \), it is necessary and sufficient that

\[
\Delta_k(m) > 0; \quad \Delta_k^{(1)}(m) > 0, \quad k \leq n
\]

and

\[
\Delta_k(m) = 0, \quad k \geq n + 1.
\]
(b) In order that there exists a solution \( \varphi \) of the moment problem (5.1) that has precisely \( n + 1 \) support points that are all in \((0, 1)\), it is necessary and sufficient that

\[
\Delta_k(m) > 0; \quad \Delta_k^{(1)}(m) > 0; \quad \Delta_k^{(2)}(m) > 0, \quad k \leq n
\]

and

\[
\Delta_k(m) = 0, \quad k \geq n + 1.
\]

(c) In order that there exists a solution \( \varphi \) of the moment problem (5.1) that has infinitely many support points that are all in \((0, \infty)\), it is necessary and sufficient that

\[
\Delta_k(m) > 0; \quad \Delta_k^{(1)}(m) > 0, \quad k = 0, 1, \ldots.
\]

(d) If there exists a solution \( \varphi \) of the moment problem (5.1) that has infinitely many support points that are all in \((0, 1)\), then

\[
\Delta_k(m) > 0; \quad \Delta_k^{(1)}(m) > 0; \quad \Delta_k^{(2)}(m) > 0, \quad k = 0, 1, \ldots.
\]

We only prove (a). Since the proofs of (b), (d) are analogous, and assertion (c) is directly from Shohat and Tamarkin (1963).

**Proof.** Necessity. Since \( \varphi \) has precisely \( n + 1 \) support points, by Lemma 5.1,

\[
\Delta_k(m) > 0, \quad k \leq n,
\]

\[
\Delta_k(m) = 0, \quad k \geq n + 1.
\]

That all \( n + 1 \) support points of \( \varphi \) are in \((0, \infty)\) implies

\[
\int_{-\infty}^{\infty} x P_n^2(x) d\varphi(x) > 0,
\]

(5.6)

where \( P_n(x) \neq 0 \) is a polynomial of degree less than or equal to \( n \), so

\[
\Delta_k^{(1)}(m) > 0, \quad k \leq n.
\]
Sufficiency. If

\[ \Delta_k(m) > 0; \quad k \leq n \]

and \[ \Delta_k(m) = 0, \quad k \geq n + 1, \]

then due to Lemma 5.1, \( \varphi \) has \( n + 1 \) support points. If \( \Delta_k^{(1)}(m) > 0, \quad k \leq n \), then (5.6) is valid for any \( P_n(x) \neq 0 \), which implies all \( n + 1 \) support points of \( \varphi \) are positive.

\[
\Box
\]

The following result is from Lindsay (1989). Lindsay proved it by using the Tchebycheff type inequality.

**Lemma 5.4.** Let \( \{F_n\} \) be a sequence of probability distribution functions, and \( F \) is a probability distribution function with precisely \( k + 1 \) distinct support points. In addition, let \( m_n = (1, m_{n1}, m_{n2}, \ldots) \) and \( m = (1, m_1, m_2, \ldots) \) be moment sequences of \( F_n \) and \( F \) respectively. If \( \Delta_{k+1}(m_n) \to 0 \) and \( m_{ni} \to m_i, 1 \leq i \leq 2k + 1 \), as \( n \to \infty \), then \( F_n \to F \) weakly.

### 5.3 Method of Moments Estimation of Mixing Distributions

#### 5.3.1 Negative Binomial Mixtures

A random variable \( X \) is said to have a negative binomial mixture distribution if

\[
p(x) := P(X = x) = \int_0^1 \left( \frac{r + x - 1}{x} \right) p^r (1 - p)^x dG(p), \quad (5.7)
\]

\( x = 0, 1, \ldots \), where \( r \) is a known positive integer and \( G(p) \) is a right continuous probability distribution function with support contained in \((0, 1)\).
Identifiability

In order to estimate the mixing distribution \( G \), we must solve the identifiability problem in mixture (5.7). That is, if there is a probability distribution function \( G_1 \) with support contained in \((0,1)\) such that

\[
\int_0^1 \left( \frac{r + x - 1}{x} \right) p^r(1-p)^xdG_1(p) = \int_0^1 \left( \frac{r + x - 1}{x} \right) p^r(1-p)^xdG(p),
\]

\( x = 0, 1, \cdots \), then can we conclude \( G_1 = G \)?

We note that if (5.8) holds, then \( m_{1i} = m_i, i = 0, 1, \cdots \), where \( m_{1i} \) and \( m_i \) are the \( i \)th moments of measures \( p^r dG_1(p) \) and \( p^r dG(p) \) respectively. Since \( p^r dG_1(p) \) and \( p^r dG(p) \) are measures that are supported on \((0,1)\), they are uniquely determined by their sequence of moments. Hence \( G_1 \equiv G \), namely, mixture (5.7) is identifiable.

Construction of a Consistent Estimate of \( G \)

Let \( G \) be the unknown true mixing distribution in (5.7). Our aim is to estimate \( G \).

Define

\[
H(q) := \int_0^q \frac{(1-t)^r}{p(0)} dG(1-t),
\]

(5.9)

where

\[
p(0) = \int_0^1 -(1-t)^r dG(1-t) = \int_0^1 p^r dG(p) > 0.
\]

Thus \( H(q) \) is a left continuous probability distribution function with support in \((0,1)\). In order to estimate \( G \), we first estimate the weighted distribution \( H \).
The $i$th moment of $H(q)$ is

\[
m_i = \int_0^1 q^i dH(q) \\
= \int_0^1 -q^i \frac{(1-q)^r}{p(0)} dG(1-q) \\
= \int_0^1 (1-p)^i \frac{p^r}{p(0)} dG(p) \\
= \frac{1}{p(0) \binom{r+i-1}{i}} \int_0^1 \binom{r+i-1}{i} p^r (1-p)^i dG(p) \\
= \frac{p(i)}{p(0) \binom{r+i-1}{i}}.
\]

Since distribution $H$ is supported on $(0,1)$, it is uniquely determined by its moment sequence $m = (m_i)_{i=0}^\infty$.

Suppose that $X_1, X_2, \cdots, X_N$ is a sample from mixture (5.7). Define

\[
\hat{p}_N(i) = \frac{1}{N} \sum_{n=1}^N 1_{\{X_n=i\}}.
\]

We presume $\hat{p}_N(0) > 0$. From the strong law of large numbers, we know that as $N \to \infty$,

\[
\hat{a}_{Ni} := \frac{\hat{p}_N(i)}{\hat{p}_N(0) \binom{r+i-1}{i}}
\]

converges to $m_i$ with probability one. We will make use of $\hat{a}_N = (\hat{a}_{Ni})_{i=0}^\infty$ to construct a probability distribution supported on $(0,1)$ which converges to $H$ weakly.

Let $k_N \leq N$ denote the nonnegative integer such that

\[
\Delta_k(\hat{a}_N) > 0, \quad \Delta^{(1)}_k(\hat{a}_N) > 0, \quad \Delta^{(2)}_k(\hat{a}_N) > 0, \quad k \leq k_N
\]

and

\[
\Delta_{k_N+1}(\hat{a}_N) \leq 0 \text{ or } \Delta^{(1)}_{k_N+1}(\hat{a}_N) \leq 0 \text{ or } \Delta^{(2)}_{k_N+1}(\hat{a}_N) \leq 0 \text{ or } k_N = N.
\]

By Lemmas 5.2 and 5.3, there is a unique left continuous probability distribution $\hat{H}_N$ with all $k_N + 1$ support points in $(0,1)$ such that

\[
\hat{m}_{Ni} := \int_0^1 q^i d\hat{H}_N(q) = \hat{a}_{Ni}, \quad i = 0, 1, \cdots, 2k_N + 1.
\]
The support points \( \hat{q}_1, \cdots, \hat{q}_{k_N+1} \) of \( \hat{H}_N \) are the solutions of
\[
\begin{pmatrix}
1 & x & \cdots & x^{k_N+1} \\
\hat{a}_{N0} & \hat{a}_{N1} & \cdots & \hat{a}_{Nk_N+1} \\
\hat{a}_{N1} & \hat{a}_{N2} & \cdots & \hat{a}_{Nk_N+2} \\
\vdots & \vdots & \ddots & \vdots \\
\hat{a}_{Nk_N} & \hat{a}_{Nk_N+1} & \cdots & \hat{a}_{N2k_N+1}
\end{pmatrix} = 0
\] (5.10)
and the corresponding weights \( \hat{w}_1, \cdots, \hat{w}_{k_N+1} \) are the solutions of
\[
\sum_{n=1}^{k_N+1} w_n \hat{q}_n^i = \hat{a}_{Ni}, \quad i = 0, 1, \cdots, k_N.
\]

We will show that \( \hat{H}_N \) converges to \( H \) weakly.

Case 1. If the support of \( H \) is not reducible to a finite set, then due to Lemma 5.3,
\[
\Delta_n(m) > 0, \quad \Delta_n^{(1)}(m) > 0, \quad \Delta_n^{(2)}(m) > 0, \quad n = 0, 1, 2, \cdots.
\]

Now we need to show that the moment sequence \( \hat{m}_N = (\hat{m}_{N0}, \hat{m}_{N1}, \cdots) \) with respect to \( \hat{H}_N \) converges almost surely to \( m \) coordinatewise, or equivalently, \( k_N \to \infty \).

Since when \( N \to \infty \), \( \hat{a}_{Ni} \to m_i \) almost surely, for any given positive integer \( K \), there exists \( n(K) \) such that when \( N \geq n(K) \),
\[
\Delta_n(\hat{a}_N) > 0, \quad \Delta_n^{(1)}(\hat{a}_N) > 0, \quad \Delta_n^{(2)}(\hat{a}_N) > 0, \quad n = 0, 1, \cdots, K,
\]
which is equivalent to
\[
k_N \geq K,
\]
which implies that
\[
k_N \to \infty \text{ as } N \to \infty.
\]

Following the Fréchet and Shohat theorem (Loève 1960), \( \hat{H}_N \to H \) weakly.

Case 2. If the support of \( H \) consists of \( K + 1 \) distinct points, then by Lemma 5.3,
\[
\Delta_n(m) > 0, \quad \Delta_n^{(1)}(m) > 0, \quad \Delta_n^{(2)}(m) > 0, \quad n = 0, 1, \cdots, K;
\]
\[
\Delta_n(m) = 0, \quad n \geq K + 1.
\]
Since $\hat{a}_{N_i} \to m_i$ almost surely, when $N$ is sufficiently large, we have

\[
\Delta_n(\hat{a}_N) > 0, \quad \Delta_n^{(1)}(\hat{a}_N) > 0, \quad \Delta_n^{(2)}(\hat{a}_N) > 0, \quad n = 0, 1, \ldots, K.
\]

Hence $k_N \geq K$. So, when $N \to \infty$,

\[
\Delta_{K+1}(\hat{m}_N) \to 0.
\]

Obviously,

\[
\hat{m}_{N_i} \to m_i, \quad i = 0, 1, \ldots, 2K + 1.
\]

Applying Lemma 5.4, we get $\hat{H}_N \to H$ weakly.

From (5.9), we obtain

\[
dG(t) = -\frac{p(0)}{tr}dH(1-t), \quad t \in (0,1),
\]

that is,

\[
\int_p^1 dG(t) = \int_p^1 -\frac{p(0)}{tr}dH(1-t)
\]

or

\[
G(p) = 1 - \int_0^{1-p} \frac{p(0)}{(1-t)r}dH(t).
\]

We can also conclude from (5.9) that if $p \in (0,1)$ is a continuous point of $G$, then $1-p$ is a continuous point of $H$, and both $p = 0$ and $p = 1$ are continuous points of $H$.

Define

\[
\hat{G}_N(p) = 1 - \int_0^{1-p} \frac{\hat{p}_N(0)}{(1-t)r}d\hat{H}_N(t), \quad p \in (0,1).
\]

Since

\[
\hat{p}_N(0)\hat{H}_N(t) \to p(0)H(t)
\]
at any continuous point \( t \) of \( H \) with probability one, as \( N \to \infty \), by the second theorem of Helly, at any continuous point \( p \in (0, 1) \) of \( G \),

\[
\tilde{G}_N(p) = 1 - \int_0^{1-p} \frac{\hat{p}_N(0)}{(1-t)^r} d\bar{H}_N(t) \to \\
G(p) = 1 - \int_0^{1-p} \frac{p(0)}{(1-t)^r} dH(t),
\]

due to \( t = 0 \) and \( t = 1 - p \) being continuous points of \( H \).

Let

\[
\tilde{G}_N(p) = \begin{cases} 
1 - \frac{1-\tilde{G}_N(p)}{\int_0^{1-p} \frac{\hat{p}_N(0)}{(1-t)^r} d\bar{H}_N(t)}, & p \in (0, 1), \\
0, & p \in (-\infty, 0], \\
1, & p \in [1, \infty).
\end{cases}
\]

Clearly, \( \tilde{G}_N(p) \) is a right continuous probability distribution function with support contained in \((0, 1)\) and \( \tilde{G}_N \) is a consistent nonparametric estimate of \( G \).

**Remark 5.1.** In the above construction of \( \hat{H}_N \), we use \( N \) to control \( k_N \) for convenience. Actually we can use a quantity larger or a smaller than \( N \) to control \( k_N \) for fitting more moments or for solving a lower degree polynomial equation (5.10) as a trade-off. Our purpose of imposing the conditions

\[
\Delta_k^{(1)}(\hat{\mathbf{a}}_N) > 0, \quad \Delta_k^{(2)}(\hat{\mathbf{a}}_N) > 0, \quad k \leq k_N
\]

is to reduce the degree of polynomial in (5.10). Without conditions

\[
\Delta_k^{(1)}(\hat{\mathbf{a}}_N) > 0, \quad \Delta_k^{(2)}(\hat{\mathbf{a}}_N) > 0, \quad k \leq k_N,
\]

we can still obtain a consistent estimate of \( H \) by truncation such that the estimator is supported on \((0, 1)\). However, sometimes we have to solve a higher degree polynomial equation (5.10).
5.3.2 Normal Mixtures

In this section, we consider a normal mixture whose density function has the form

\[ f(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dG(\mu), \quad -\infty < x < \infty, \]

(5.11)

where \( \sigma > 0 \) is known, and \( G(\mu) \) is a probability distribution function on \( (-\infty, \infty) \).

Identifiability

Let \( G_1, G_2 \) be two probability distribution functions on \( (-\infty, \infty) \). If

\[
\int_{-\infty}^{\infty} e^{itx} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dG_1(\mu)dx \\
= \int_{-\infty}^{\infty} e^{itx} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dG_2(\mu)dx.
\]

After simplification, we obtain

\[
\int_{-\infty}^{\infty} e^{it\mu} dG_1(\mu) = \int_{-\infty}^{\infty} e^{it\mu} dG_2(\mu).
\]

That is, the characteristic functions of \( G_1, G_2 \) are same, so \( G_1 \equiv G_2 \).

Constructing a Consistent Nonparametric Estimate of \( G \)

We presume that the mixing distribution \( G \) in (5.11) satisfies the condition

\[
\int_{-\infty}^{\infty} e^{c|x|} dG(\mu) < \infty, \quad \text{for some } r > 0.
\]

(5.12)

for some \( r > 0 \). Usually \( G(\mu) \) is concentrated on a finite interval, or it assigns little density or mass to \( \mu \) with very large absolute value. So condition (5.12) is often satisfied.
Let $X$ be a random variable with normal mixture density $f$ in (5.11). Now we calculate $E(e^{iX - \frac{z^2 \sigma^2}{2}})$, where $i$ is a nonnegative integer.

\[
E(e^{iX - \frac{z^2 \sigma^2}{2}}) = \int_{-\infty}^{\infty} e^{i(x - \frac{z^2 \sigma^2}{2})} f(x) dx
\]
\[
= \int_{-\infty}^{\infty} e^{i(x - \frac{z^2 \sigma^2}{2})} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(x - \mu)^2}{2\sigma^2}} dG(\mu) dx
\]
\[
= \int_{-\infty}^{\infty} e^{i\mu} dG(\mu).
\]

The last integral is finite, since when $\mu$ is large enough, $i\mu < e^{\nu r}$. In order to use the moments of $e^\mu$, we define

\[
H(y) = \begin{cases} 
G(\ln y), & y > 0, \\
0, & y \leq 0.
\end{cases}
\]

(5.13)

Here $H$ is a probability distribution function supported on $(0, \infty)$ with the $i$th-moment

\[
m_i := \int_{0}^{\infty} y^i dH(y) = E(e^{iX - \frac{z^2 \sigma^2}{2}}).
\]

Since for any $n$,

\[
\sum_{i=0}^{n} \frac{\int_{-\infty}^{\infty} e^{i\mu} dG(\mu)}{i!} r^i = \int_{-\infty}^{\infty} \sum_{i=0}^{n} \frac{e^{i\mu r^i}}{i!} dG(\mu)
\]
\[
\leq \int_{-\infty}^{\infty} \sum_{i=0}^{\infty} \frac{e^{i\mu r^i}}{i!} dG(\mu)
\]
\[
= \int_{-\infty}^{\infty} e^{e^{\nu r}} dG(\mu),
\]

\[
\sum_{i=0}^{\infty} \frac{m_i r^i}{i!} = \sum_{i=0}^{\infty} \frac{\int_{-\infty}^{\infty} e^{i\mu} dG(\mu)}{i!} r^i
\]
\[
\leq \int_{-\infty}^{\infty} e^{e^{\nu r}} dG(\mu) < \infty.
\]

By a result in Cramér (1946, page 176), $H$ is uniquely determined by its moment sequence $m = (m_i)_{i=0}^{\infty}$. 
Suppose \( X_1, \cdots, X_N \) is a sample from normal mixture (5.11). According to the strong law of large numbers,

\[
\hat{\alpha}_{Ni} := \frac{1}{N} \sum_{j=1}^{N} e^{iX_j - \frac{a_{Ni}^2}{2}} \to m_i
\]

almost surely, when \( N \) tends to infinity. We want to modify \( \hat{\alpha}_N = (\hat{\alpha}_{Ni})_{i=0}^{\infty} \) into a moment sequence such that the underlying distribution function is supported on \((0, \infty)\) and converges to \( H \) at all continuous points of \( H \).

Let \( k_N \leq N \) denote the nonnegative integer such that

\[
\Delta_k(\hat{\alpha}_N) > 0, \quad \Delta_k^{(1)}(\hat{\alpha}_N) > 0, \quad k \leq k_N
\]

and

\[
\Delta_{kN+1}(\hat{\alpha}_N) \leq 0 \text{ or } \Delta_{kN+1}^{(1)}(\hat{\alpha}_N) \leq 0 \text{ or } k_N = N.
\]

Following Lemmas 5.2 and 5.3, there exists a unique probability distribution function \( \hat{H}_N \) with all \( k_N + 1 \) support points in \((0, \infty)\) fitting the first \( 2k_N + 2 \) moments \( \hat{\alpha}_{N0}, \cdots, \hat{\alpha}_{N2k_N+1} \). That is,

\[
\hat{m}_{Ni} := \int_0^\infty y^i d\hat{H}_N(y) = \hat{\alpha}_{Ni}, \quad i = 0, 1, \cdots, 2k_N + 1.
\]

In a similar manner to that in section 5.3.1, we can prove that \( \hat{H}_N \) converges to \( H \) at every continuous point of \( H \).

Define

\[
\hat{G}_N(\mu) = \hat{H}_N(e^\mu).
\]

Suppose \( \mu \) is a continuous point of \( G \). Then from (5.9), \( e^\mu \) is a continuous point of \( H \). Thus

\[
\hat{H}_N(e^\mu) \to H(e^\mu),
\]

or equivalently,

\[
\hat{G}_N(\mu) \to G(\mu).
\]

Therefore, \( \hat{G}_N \) is a consistent estimate of \( G \).
5.3.3 Binomial Mixtures

In this section, we consider the following binomial mixture

\[ p(x) := P(X = x) = \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} dG(p), \]  

(5.14)

\( x = 0, 1, \ldots, n \), where \( n \) is a fixed positive integer and \( G(p) \) is a probability distribution function supported on \((0,1)\) with at most \( \frac{n+1}{2} \) support points.

By Proposition 4 in Teicher (1963), the mixture (5.14) is identifiable. Clearly, \( G \) is uniquely determined by its first \( n + 1 \) moments \( (m_k)_{k=0}^n \). From (5.14), we have

\[ \begin{align*}
    p(n-i) &= \int_0^1 \binom{n}{n-i} p^{n-i} (1-p)^i dG(p) \\
    &= \int_0^1 \binom{n}{n-i} p^{n-i} \sum_{j=0}^{i} \binom{i}{j} (-1)^{i-j} p^{-j} dG(p) \\
    &= \binom{n}{n-i} \sum_{j=0}^{i} \binom{i}{j} (-1)^{i-j} m_{n-j},
\end{align*} \]

\( i = 0, 1, \ldots, n \).

Thus

\[ m_{n-i} = \sum_{j=0}^{i} \binom{n}{n-j}^{-1} \binom{i}{j} p(n-j), \quad i = 0, 1, \ldots, n. \]

Suppose that \( X_1, X_2, \ldots, X_N \) is a sample from the mixture (5.14). Define

\[ \hat{p}_N(n-j) = \frac{1}{N} \sum_{k=1}^{N} 1\{X_k = n-j\} \]

and

\[ \hat{a}_{Nn-i} = \sum_{j=0}^{i} \binom{n}{n-j}^{-1} \binom{i}{j} \hat{p}_N(n-j). \]

By the strong law of large numbers, when \( N \to \infty \),

\[ \hat{a}_{Ni} \to m_i, \quad i = 0, 1, \ldots, n. \]
Denote $\hat{a}_N = (\hat{a}_{Ni})_{i=0}^n$. Let $k_N \leq \left\lfloor \frac{n-1}{2} \right\rfloor$, the greatest integer less than or equal to $\frac{n-1}{2}$, denote the nonnegative integer such that

$$\Delta_k(\hat{a}_N) > 0, \quad \Delta_k^{(1)}(\hat{a}_N) > 0, \quad \Delta_k^{(2)}(\hat{a}_N) > 0, \quad k \leq k_N$$

and

$$\Delta_{kN+1}(\hat{a}_N) \leq 0 \text{ or } \Delta_{kN+1}^{(1)}(\hat{a}_N) \leq 0 \text{ or } \Delta_{kN+1}^{(2)}(\hat{a}_N) \leq 0 \text{ or } k_N = \left\lfloor \frac{n-1}{2} \right\rfloor.$$

Following Lemmas 5.2 and 5.3, there is a unique probability distribution function $\hat{G}_N$ with all $k_N + 1$ support points in $(0,1)$ such that

$$\hat{m}_{Ni} := \int_0^1 p^i d\hat{G}_N(p) = \hat{a}_{Ni}, \quad i = 0, 1, \ldots, 2k_N + 1.$$

Define

$$\hat{m}_{Ni} := \int_0^1 p^i d\hat{G}_N(p), \quad i = 0, 1, \ldots$$

and denote $\hat{\mathbf{m}}_N = (\hat{m}_{Ni})_{i=0}^\infty$. If the support of $G$ consists of $K + 1 \leq \frac{n+1}{2}$ distinct points, then

$$\Delta_k(\mathbf{m}) > 0, \quad \Delta_k^{(1)}(\mathbf{m}) > 0, \quad \Delta_k^{(2)}(\mathbf{m}) > 0, \quad k \leq K;$$

and

$$\Delta_k(\mathbf{m}) = 0, \quad k \geq K + 1,$$

where $\mathbf{m}$ is the moment sequence of $G$. Since as $N \to \infty$, with probability one

$$\hat{a}_{Ni} \to m_i, \quad i = 0, 1, \ldots, n,$$

when $N$ is sufficiently large, we have

$$\Delta_k(\hat{a}_N) > 0, \quad \Delta_k^{(1)}(\hat{a}_N) > 0, \quad \Delta_k^{(2)}(\hat{a}_N) > 0, \quad k \leq K.$$

Hence for enough large $N$, $k_N \geq K$. Thus we have

$$\Delta_{K+1}(\hat{m}_N) \to 0$$
and

\[ \hat{m}_{N_i} \to m_i, \quad i = 0, 1, \ldots, 2K + 1. \]

Following Lemma 5.4,

\[ \hat{G}_N \to G \]

weakly. Therefore we obtain a consistent nonparametric estimate \( \hat{G}_N \) of \( G \).
Chapter 6

Method of Moments Estimation in Finite Mixture Models

6.1 Introduction

In this chapter, we focus on the estimation of the mixing distribution $G$ in a finite mixture

$$f(x; G) = \sum_{i=1}^{n+1} \pi_i f(x; \lambda_i),$$

(6.1)

where $n \geq 1$ is a fixed integer, $f(x; \lambda)$, $\lambda \in [0, 1]$ is a parametric family and

$$G = \sum_{i=1}^{n+1} \pi_i \delta(\lambda_i)$$

is an $(n+1)$-point distribution on $[0, 1]$.

Lindsay (1989) studied the method of moments in finite mixtures where the number of support points of the mixing distribution is fixed. He considered several cases in which the estimated distribution violates the constraints on the parameter space, and proposed a method, without proof, to correct the distribution function so that its support is within the specified parameter space. His modification method not only adjusts the estimator while matching some initial moments, but also reduces the polynomial
equation by one or two degrees to obtain the support points of the estimator. This reduces the computational complexity of the problem.

In Section 6.2, we provide a preliminary lemma that allows us in Section 6.3 to prove Lindsay’s (1989) modification result, as well as the weak consistency of the resulting estimator. In Section 6.4, we establish the convergence rate of the modified estimator in this situation as $N^{-\frac{1}{2}} (\log \log N)^{\frac{1}{2}}$.

### 6.2 A Preliminary Lemma

**Lemma 6.1.** Let $\varphi(x)$ be a distribution function with exactly $n + 1$ support points $x_1 < x_2 < \cdots < x_{n+1}$ and let $m = (m_0, m_1, \cdots)$ be its moment sequence. Then

(a) 
\[
\Delta_k^{(1)}(m) > 0; \quad \Delta_k^{(2)}(m) > 0, \quad k \leq n
\]

if and only if
\[
x_1 < x_2 < \cdots < x_{n+1} \in (0, 1).
\]

(b) If
\[
\Delta_{k-1}^{(1)}(m) > 0, \quad \Delta_{n}^{(1)}(m) \leq 0; \quad \Delta_k^{(2)}(m) > 0, \quad k \leq n,
\]

then
\[
x_1 \leq 0, \quad x_2 < \cdots < x_{n+1} \in (0, 1).
\]

(c) If
\[
\Delta_k^{(1)}(m) > 0; \quad \Delta_k^{(2)}(m) > 0, \quad \Delta_{k-1}^{(2)}(m) \leq 0, \quad k \leq n,
\]

then
\[
x_{n+1} \geq 1, \quad x_1 < \cdots < x_n \in (0, 1).
\]
(d) If

\[ \Delta_k^{(1)}(m) > 0, \quad \Delta_n^{(1)}(m) \leq 0; \quad \Delta_k^{(2)}(m) > 0, \quad \Delta_n^{(2)}(m) \leq 0, \quad k \leq n - 1, \]

then

\[ x_1 \leq 0, \quad x_{n+1} \geq 1, \quad x_2 < \cdots < x_n \in (0,1). \]

**Proof.** For (a), consider the integral

\[ \int_{-\infty}^{\infty} x P_n^2(x) d\varphi(x), \]

where \( P_n(x) \) is a real polynomial of degree less than or equal to \( n \). If

\[ \Delta_k^{(1)}(m) > 0, \quad k \leq n, \]

then

\[ \int_{-\infty}^{\infty} x P_n^2(x) d\varphi(x) > 0 \]

for any \( P_n(x) \neq 0 \). For any \( 1 \leq i \leq n+1 \), choose a special \( P_n(x) \) such that

\[ x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{n+1} \]

are the roots of \( P_n(x) \). If we assume that \( w_i \) is the weight of \( x_i \), then

\[ \int_{-\infty}^{\infty} x P_n^2(x) d\varphi(x) = x_i w_i P_n^2(x_i) > 0, \]

which implies \( x_i > 0 \). Conversely, if \( x_i > 0, i = 1, \cdots, n+1 \), then

\[ \int_{-\infty}^{\infty} x P_n^2(x) d\varphi(x) > 0 \]

for any \( P_n(x) \neq 0 \). So

\[ \Delta_k^{(1)}(m) > 0, \quad k \leq n. \]
Now consider the integral

$$\int_{-\infty}^{\infty} (1 - x)P_n^2(x) d\varphi(x),$$

where $P_n(x)$ is a real polynomial of degree less than or equal to $n$. If

$$\Delta_k^{(2)}(m) > 0, \quad k \leq n,$$

then

$$\int_{-\infty}^{\infty} (1 - x)P_n^2(x) d\varphi(x) > 0$$

for any $P_n(x) \neq 0$. For any $1 \leq i \leq n + 1$, if we choose a special $P_n(x)$ such that

$x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{n+1}$ are the roots of $P_n(x)$, then

$$\int_{-\infty}^{\infty} (1 - x)P_n^2(x) d\varphi(x) = (1 - x_i)w_i P_n^2(x_i) > 0,$$

which implies $1 - x_i > 0$ or $x_i < 1$. Conversely, if $x_i < 1, i = 1, \cdots, n + 1$, then

$$\int_{-\infty}^{\infty} (1 - x)P_n^2(x) d\varphi(x) > 0$$

for any $P_n(x) \neq 0$. So

$$\Delta_k^{(2)}(m) > 0, \quad k \leq n.$$

For (b), from the proof of (a), we know $x_1 < 1, x_1, \ldots$, $x_{n+1}$. Since

$$\Delta_n^{(1)}(m) \leq 0,$$

$x_1 \leq 0$. If $x_2 \leq 0$, by choosing an appropriate real polynomial $P_{n-1}(x)$ of degree $n - 1$

such that $P_{n-1}(x_i) = 0, 3 \leq i \leq n + 1$, then we have

$$\int_{-\infty}^{\infty} xP_{n-1}^2(x) d\varphi(x) = x_1w_1 P_{n-1}^2(x_1) + x_2w_2 P_{n-1}^2(x_2) \leq 0,$$

which contradicts

$$\Delta_k^{(1)}(m) > 0, \quad k \leq n - 1.$$

So $x_2 > 0$. Thus, we can conclude that $x_1 \leq 0, x_2 < \cdots < x_{n+1} \in (0, 1)$.

Finally, note that the proofs for (c) and (d) are similar to the proof for (b). □
6.3 Estimation and Modification

Lindsay (1989) has shown that if \( f(x; \lambda) \) is a member of the quadratic variance natural exponential family (QVEF) (the variance is at most a quadratic function of the mean), and is parameterized in its mean value, then the estimator \( \hat{a}_{Ni} \) of the \( i \)-th moment \( m_i \) of the mixing distribution \( G \) is easily found such that

\[
\hat{a}_{Ni} \to m_i
\]

almost surely as the sample size \( N \to \infty \).

Morris (1982, 1983) demonstrated that the normal, Poisson, gamma, binomial, negative binomial, and NEF-GHS (natural exponential family generated by the hyperbolic secant) distributions are the six univariate natural exponential families (NEF) with quadratic variance functions (QVF). Throughout this paper, we assume \( \{f(x; \lambda), \lambda \in [0, 1]\} \) is a QVEF and \( \lambda \) is the mean value. If \( a(f) > 0, \ k < n \)

which is always true when \( N \) is large enough, then there exists an \( (n + 1) \)-point distribution \( \hat{G}_N \) fitting moments \( \hat{a}_{N0}, \hat{a}_{N1}, \ldots, \hat{a}_{N2n+1} \). This distribution \( \hat{G}_N \) converges weakly to \( G \).

Suppose that in model (6.1) the true mixing distribution \( G \) has \( \lambda = 0 \) or \( \lambda = 1 \) as one of its support points. Then, provided that \( N \) is sufficiently large, with positive probability at least one of support points of the estimated distribution \( \hat{G}_N \) lies outside of \( [0, 1] \). Lindsay (1989) proposed a modification such that the corrected estimator \( \hat{G}_N \) fits the initial \( 2n + 1 \) or \( 2n \) moments. Here we give the proof of this modification as well as the weak consistency of the estimator.

Let \( t_1 < t_2 < \cdots < t_{n+1} \) be the support points of \( \hat{G}_N \). According to Lemma 6.1
(i) If
\[ \Delta_k^{(1)}(\hat{a}_N) > 0; \quad \Delta_k^{(2)}(\hat{a}_N) > 0, \quad k \leq n, \]
then
\[ t_1 < t_2 < \cdots < t_{n+1} \in (0,1). \]

(ii) If
\[ \Delta_{k-1}^{(1)}(\hat{a}_N) > 0, \quad \Delta_n^{(1)}(\hat{a}_N) \leq 0; \quad \Delta_k^{(2)}(\hat{a}_N) > 0, \quad k \leq n, \]
then
\[ t_1 \leq 0, \quad t_2 < \cdots < t_{n+1} \in (0,1). \]

(iii) If
\[ \Delta_k^{(1)}(\hat{a}_N) > 0; \quad \Delta_{k-1}^{(2)}(\hat{a}_N) > 0, \quad \Delta_k^{(2)}(\hat{a}_N) \leq 0, \quad k \leq n, \]
then
\[ t_{n+1} \geq 1, \quad t_1 < \cdots < t_n \in (0,1). \]

(iv) If
\[ \Delta_k^{(1)}(\hat{a}_N) > 0, \quad \Delta_n^{(1)}(\hat{a}_N) \leq 0; \quad \Delta_k^{(2)}(\hat{a}_N) > 0, \quad \Delta_n^{(2)}(\hat{a}_N) \leq 0, \quad k \leq n-1, \]
then
\[ t_1 \leq 0, \quad t_{n+1} \geq 1, \quad t_2 < \cdots < t_n \in (0,1). \]

For case (i), no modification is needed. For case (ii), an \((n+1)\)-point modified distribution \(\tilde{G}_N\) with \(t_1 = 0\) that fits the first \(2n+1\) moments \(\hat{a}_{N0}, \hat{a}_{N1}, \ldots, \hat{a}_{N2n}\) is derived as follows. If \(t_2, \cdots, t_{n+1}\) are the \(n\) support points of \(\tilde{G}_N\) other than zero, and \(p_2, \cdots, p_{n+1}\) are the corresponding weights, then
\[ \hat{a}_{Ni} = \sum_{j=2}^{n+1} p_j t_j^i, \quad i = 1, \cdots, 2n. \quad (6.2) \]
Define
\[ \hat{m}_N^i := \hat{a}_N^i, \quad i \leq 2n, \]
\[ \hat{m}_N^i := \sum_{j=2}^{n+1} p_j t_j^i, \quad i \geq 2n + 1. \]

Then \( \hat{m}_N = (\hat{m}_N^i)_{i=0}^\infty \) is the moment sequence of this distribution. Since \( t = 0 \) is a solution of
\[
\begin{bmatrix}
1 & t & \cdots & t^{n+1} \\
\hat{m}_N^0 & \hat{m}_N^1 & \cdots & \hat{m}_N^{n+1} \\
\hat{m}_N^1 & \hat{m}_N^2 & \cdots & \hat{m}_N^{n+2} \\
\vdots & \vdots & \ddots & \vdots \\
\hat{m}_N^n & \hat{m}_N^{n+1} & \cdots & \hat{m}_N^{2n+1}
\end{bmatrix} = 0,
\]
we have
\[ \Delta_n^{(1)}(\hat{m}_N) = 0. \]

Following Theorem 1.3 in Shohat and Tamarkin (1963), \( t_2, \ldots, t_{n+1} > 0 \).

If in (6.2) we take \( p_j t_j > 0 \) as the weight of \( t_j \) and consider the \( n \)-point distribution that has its first \( 2n \) moments \( \hat{a}_N^i, \ i = 1, \ldots, 2n \) (where \( \hat{a}_N^i \) is the \( (i-1) \)-th moment), then by Lemma 5.2, all \( t_j, \ 2 \leq j \leq n + 1 \), are solutions of
\[
\begin{bmatrix}
1 & t & \cdots & t^n \\
\hat{a}_N^1 & \hat{a}_N^2 & \cdots & \hat{a}_N^{n+1} \\
\hat{a}_N^2 & \hat{a}_N^3 & \cdots & \hat{a}_N^{n+2} \\
\vdots & \vdots & \ddots & \vdots \\
\hat{a}_N^n & \hat{a}_N^{n+1} & \cdots & \hat{a}_N^{2n}
\end{bmatrix} = 0. \quad (6.3)
\]

For case (iii), we create a distribution \( \tilde{G}_N \) with \( t_{n+1} = 1 \) that fits the initial \( 2n + 1 \) moments \( \hat{a}_N^0, \hat{a}_N^1, \cdots, \hat{a}_N^{2n} \). If we let \( t_1, \cdots, t_n \) be the \( n \) support points of \( \tilde{G}_N \) other than one, then
\[ \hat{a}_N^i = \sum_{j=1}^n p_j t_j^i + p_{n+1}, \quad i = 0, 1, \cdots, 2n. \]

Suppose \( \tilde{m}_N = (\tilde{m}_N^i)_{i=0}^\infty \) is the moment sequence of this distribution. Because
\[ \Delta_k^{(2)}(\tilde{m}_N) > 0, \quad k \leq n - 1, \]
by the theory of quadratic form,

\[ \int_{-\infty}^{\infty} (1 - t)P_{n-1}^2(t) d\tilde{G}_N(t) > 0 \]

for any less than or equal to \((n - 1)\) degree polynomial \(P_{n-1}(t) \neq 0\), which implies that \(t_1, \ldots, t_n < 1\). Since

\[
\hat{a}_{N_1} - \hat{a}_{N_1+1} = \sum_{j=1}^{n} p_j (1 - t_j) t_j^i, \quad i = 0, 1, \ldots, 2n - 1,
\]

if we take \(p_j (1 - t_j) > 0\) as the weight of \(t_j\) and consider an \(n\)-point distribution that fits the initial \(2n\) moments \(\hat{a}_{N_1} - \hat{a}_{N_1+1}, i = 0, 1, \ldots, 2n - 1\), then by Lemma 5.2, all \(t_j, 1 \leq j \leq n\) are the solutions of

\[
\begin{vmatrix}
1 & t & \cdots & t^n \\
\hat{a}_{N_0} - \hat{a}_{N_1} & \hat{a}_{N_1} - \hat{a}_{N_2} & \cdots & \hat{a}_{N_n} - \hat{a}_{N_{n+1}} \\
\hat{a}_{N_1} - \hat{a}_{N_2} & \hat{a}_{N_2} - \hat{a}_{N_3} & \cdots & \hat{a}_{N_{n+1}} - \hat{a}_{N_{n+2}} \\
\vdots & \vdots & \vdots & \vdots \\
\hat{a}_{N_{n-1}} - \hat{a}_{N_n} & \hat{a}_{N_n} - \hat{a}_{N_{n+1}} & \cdots & \hat{a}_{N_{2n-1}} - \hat{a}_{N_{2n}}
\end{vmatrix} = 0. \quad (6.4)
\]

If there is a support point derived above in case (ii) that is larger than 1, or there is a support point derived above in case (iii) that is less than zero, or case (iv) occurs, we only fit the first \(2n\) moments \(\hat{a}_{N_0}, \hat{a}_{N_1}, \ldots, \hat{a}_{N_{2n-1}}\). We create a modified distribution \(\tilde{G}_N\) with \(t_1 = 0, t_{n+1} = 1\). Let \(t_2, \ldots, t_n\) be the \(n - 1\) support points of \(\tilde{G}_N\) that are neither zero nor one. Then

\[
\hat{a}_{Ni} = \sum_{j=2}^{n} p_j t_j^i + p_{n+1}, \quad i = 1, \ldots, 2n - 1.
\]

Suppose \(\hat{m}_N = (\hat{m}_{N_i})_{i=0}^{\infty}\) is the moment sequence of this distribution. Since \(t = 0, t = 1\) are the solutions of

\[
\begin{vmatrix}
1 & t & \cdots & t^{n+1} \\
\hat{m}_{N_0} & \hat{m}_{N_1} & \cdots & \hat{m}_{N_{n+1}} \\
\hat{m}_{N_1} & \hat{m}_{N_2} & \cdots & \hat{m}_{N_{n+2}} \\
\vdots & \vdots & \vdots & \vdots \\
\hat{m}_{N_n} & \hat{m}_{N_{n+1}} & \cdots & \hat{m}_{N_{2n+1}}
\end{vmatrix}
\]
we have

\[ \Delta_n^{(1)}(\hat{m}_N) = 0, \quad \Delta_n^{(2)}(\hat{m}_N) = 0. \]

Following Lemma 6.1 (d), \( t_2, \ldots, t_n \in (0, 1) \). Since

\[ \hat{a}_{N_i} - \hat{a}_{N_{i+1}} = \sum_{j=2}^{n} p_j (1 - t_j) t_j^{i-1}, \quad i = 1, 2, \ldots, 2n - 2, \]

if we take \( p_j (1 - t_j) t_j > 0 \) as the weight of \( t_j \) and consider the \((n-1)\)-point distribution that has the first \( 2n - 2 \) moments \( \hat{a}_{N_i} - \hat{a}_{N_{i+1}}, i = 1, \ldots, 2n - 2 \) (where \( \hat{a}_{N_i} - \hat{a}_{N_{i+1}} \) is the \((i-1)\)-th moment), then by Lemma 5.2, all \( t_j, 2 \leq j \leq n \) are the solutions of

\[
\begin{vmatrix}
1 & t & \cdots & t^{n-1} \\
\hat{a}_{N_1} - \hat{a}_{N_2} & \hat{a}_{N_2} - \hat{a}_{N_3} & \cdots & \hat{a}_{N_{n-1}} - \hat{a}_{N_n} \\
\hat{a}_{N_2} - \hat{a}_{N_3} & \hat{a}_{N_3} - \hat{a}_{N_4} & \cdots & \hat{a}_{N_{n-2}} - \hat{a}_{N_{n-1}} \\
\vdots & \vdots & \ddots & \vdots \\
\hat{a}_{N_{n-1}} - \hat{a}_{N_n} & \hat{a}_{N_n} - \hat{a}_{N_{n+1}} & \cdots & \hat{a}_{N_{2n-2}} - \hat{a}_{N_{2n-1}} \\
\end{vmatrix} = 0. \tag{6.5}
\]

**Remark 6.1.** In fact, when case (ii), (iii) or (iv) occurs, we can simply truncate \( \hat{G}_N \) to get

\[
G^*_N(t) = \begin{cases} 
\hat{G}_N(t), & 0 \leq t < 1, \\
0, & t < 0, \\
1, & t \geq 1.
\end{cases}
\]

Clearly, \( G^*_N \) also converges weakly to \( G \). However, as Lindsay (1989) pointed out, \( G^*_N \) no longer fits any of the moments. Moreover, in order to obtain \( G^*_N \), we must solve the
to obtain the \( n + 1 \) support points of \( \tilde{G}_N \), then process truncation. In Lindsay’s (1989) modification, in order to obtain the support points of \( \tilde{G}_N \), we only need solve the polynomial equation (6.3), (6.4), or (6.5) which are one or two degrees less than equation (6.6), thus reducing computational complexity. Following Lemma 5.2, the corresponding weights can be calculated by solving a system of linear equations.

**Theorem 6.1.** The above modified estimator \( \tilde{G}_N \) is weakly consistent.

**Proof.** 1. If \( 0 < \lambda_1 < \lambda_2 < \cdots < \lambda_{n+1} < 1 \), then when \( N \) is sufficiently large, only case (i) happens. The support points of \( \tilde{G}_N \) (same as \( \hat{G}_N \) in this case) are the roots of (6.6), and the weak convergence of \( \tilde{G}_N \) to \( G \) is obvious, since the support points of \( G \) satisfy (6.6) with \( \hat{a}_N \) replaced by \( m \).

2. If \( 0 = \lambda_1 < \lambda_2 < \cdots < \lambda_{n+1} < 1 \), then only case (i) or case (ii) can occur when \( N \) is large enough. If case (i) occurs, then the situation is the same as when \( 0 < \lambda_1 < \lambda_2 < \cdots < \lambda_{n+1} < 1 \). If case (ii) occurs, then one support point of \( \tilde{G}_N \) is 0, and all the other support points of \( \tilde{G}_N \) are the roots of (6.3). The mixing distribution \( G \) has the same property, with \( \hat{a}_N \) replaced by \( m \) in (6.3). Thus, regardless of whether case (i) or case (ii) occurs, the estimator \( \tilde{G}_N \) converges weakly to \( G \).

3. If \( 0 < \lambda_1 < \lambda_2 < \cdots < \lambda_{n+1} = 1 \), then only case (i) or case (iii) can occur when \( N \) is sufficiently large. The analogous analysis to that for the situation when \( 0 = \lambda_1 < \lambda_2 < \cdots < \lambda_{n+1} < 1 \) demonstrates the weak convergence of \( \tilde{G}_N \).

4. If \( 0 = \lambda_1 < \lambda_2 < \cdots < \lambda_{n+1} = 1 \), then any of the cases from (i) to (iv) can

\[
\begin{array}{cccc}
1 & t & \cdots & t^{n+1} \\
\hat{a}_{N0} & \hat{a}_{N1} & \cdots & \hat{a}_{Nn+1} \\
\hat{a}_{N1} & \hat{a}_{N2} & \cdots & \hat{a}_{Nn+2} \\
\vdots & \vdots & \ddots & \vdots \\
\hat{a}_{Nn} & \hat{a}_{Nn+1} & \cdots & \hat{a}_{N2n+1} \\
\end{array} = 0
\] (6.6)
occur when $N$ is large enough. If case (i) or case (ii) occurs, the analogous analysis to that for the situation when $0 = \lambda_1 < \lambda_2 < \cdots < \lambda_{n+1} < 1$ demonstrates the weak convergence of $\tilde{G}_N$. If case (iii) occurs, then one support point of $\tilde{G}_N$ is 1, and the remaining support points are the roots of (6.4). The mixing distribution $G$ has the same property, but with $\hat{a}_N$ replaced by $m$ in (6.4). If case (iv) happens, then both 0 and 1 are support points of $\tilde{G}_N$ and the remaining support points are the roots of (6.5). The mixing distribution $G$ has the same property, but with $\hat{a}_N$ replaced by $m$ in (6.5).

Hence, regardless of which of the cases (i) through (iv) occurs, $\tilde{G}_N$ is consistent.

\[ \square \]

6.4 Convergence Rate

In this section, we will show that the convergence rate of $\tilde{G}_N$ is $N^{-\frac{1}{2}}(\log \log N)^{\frac{1}{2}}$. Note that Chen (1995) established the best possible rate $N^{-\frac{1}{2}}$ of convergence associated with the maximum likelihood estimate in finite mixture models when the exact number of support points of the mixing distribution is unknown, while the $N^{-\frac{1}{2}}$ convergence rate is achievable when the exact number of support points is known. He used the quantity

$$d(G_1, G_2) = \int_{\Lambda} |G_1(\lambda) - G_2(\lambda)| d\lambda$$

to measure the discrepancy of the two mixing distributions $G_1$ and $G_2$, where $\Lambda$ is a compact set. We adopt the same measure.

**Theorem 6.2.** The convergence rate of $\tilde{G}_N$ is $N^{-\frac{1}{2}}(\log \log N)^{\frac{1}{2}}$.

**Proof.** To begin with, assume that

$$0 < \lambda_1 < \lambda_2 < \cdots < \lambda_{n+1} < 1.$$
When $N$ is sufficiently large, only case (i) occurs. The estimator $\hat{G}_N$ matches the first $2n + 2$ moments

$$\hat{a}_{N0}, \hat{a}_{N1}, \ldots, \hat{a}_{N2n+1}.$$ 

Let

$$0 < t_1 < t_2 < \cdots < t_{n+1} < 1$$

be the support points of $\hat{G}_N$ and

$$p_1, p_2, \cdots, p_{n+1}$$

be the respective masses. Then

$$\hat{a}_{N0} - m_0 = \sum_{j=1}^{n+1} (p_j - \pi_j),$$

$$\hat{a}_{Ni} - m_i = \sum_{j=1}^{n+1} (p_j t_j^i - \pi_j \lambda_j^i)$$

$$= \sum_{j=1}^{n+1} \left[ \left( \sum_{k=0}^{i-1} p_j \lambda_j^k t_j^{i-1-k} \right) (t_j - \lambda_j) + \lambda_j^i (p_j - \pi_j) \right], \quad i = 1, \ldots, 2n + 1.$$

These moment equations can be written in matrix form

\[
\begin{pmatrix}
\hat{a}_{N0} - m_0 \\
\hat{a}_{N1} - m_1 \\
\vdots \\
\hat{a}_{N2n+1} - m_{2n+1}
\end{pmatrix}
= A_N
\begin{pmatrix}
t_1 - \lambda_1 \\
t_2 - \lambda_2 \\
\vdots \\
t_{n+1} - \lambda_{n+1}
\end{pmatrix}.
\]

(6.7)

If we denote the limiting matrix of $A_N$ by $A$, then

\[
A = \begin{pmatrix}
0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 \\
\pi_1 & \pi_2 & \cdots & \pi_{n+1} & \lambda_1 & \lambda_2 & \cdots & \lambda_{n+1} \\
2\pi_1 \lambda_1 & 2\pi_2 \lambda_2 & \cdots & 2\pi_{n+1} \lambda_{n+1} & \lambda_1^2 & \lambda_2^2 & \cdots & \lambda_{n+1}^2 \\
\vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\
(2n + 1)\pi_1 \lambda_1^{2n} & (2n + 1)\pi_2 \lambda_2^{2n} & \cdots & (2n + 1)\pi_{n+1} \lambda_{n+1}^{2n} & \lambda_1^{2n+1} & \lambda_2^{2n+1} & \cdots & \lambda_{n+1}^{2n+1}
\end{pmatrix}.
\]

We show that matrix $A$ is invertible in Appendix 1. Thus, when $N$ is sufficiently large, $A_N$ is also invertible. By (6.7)
\[
\begin{pmatrix}
  t_1 - \lambda_1 \\
  t_2 - \lambda_2 \\
  \vdots \\
  t_{n+1} - \lambda_{n+1} \\
  p_1 - \pi_1 \\
  \vdots \\
  p_{n+1} - \pi_{n+1}
\end{pmatrix}
= A_{N}^{-1}
\begin{pmatrix}
  \hat{a}_{N0} - m_0 \\
  \hat{a}_{N1} - m_1 \\
  \vdots \\
  \hat{a}_{N2n+1} - m_{2n+1}
\end{pmatrix}.
\]

(6.8)

Suppose \( f(x; \lambda) \) is the density function of a random variable \( Y \). According to Morris (1982), there is a polynomial \( \gamma_i(Y) \) of degree \( i \) in \( Y \) that is an unbiased estimator of \( \lambda^i \). Further let \( X \) be the random variable having mixture density \( f(x; G) \). Then \( \gamma_i(X) \) is an unbiased estimator of the \( i \)-th moment \( m_i \) of \( G \). Therefore we may take \( \hat{a}_{Ni} \) as

\[
\frac{\sum_{k=1}^{N} \gamma_i(X_k)}{N}
\]

for a sample \( X_1, \cdots, X_N \) from \( f(x; G) \). By the strong law of large numbers,

\[
\hat{a}_{Ni} \to m_i
\]

almost surely.

Suppose that \( \sigma_i^2 \) is the variance of \( \gamma_i(X) \). Then, appealing to the law of the iterated logarithm of Hartman-Wintner (1941), which asserts that with probability one the set of limit points of the sequence

\[
\left( (2N \log \log N)^{\frac{1}{3}} \sum_{k=1}^{N} \frac{\gamma_i(X_k) - m_i}{\sigma_i} \right)_{N \geq 3}
\]

coincides with \([-1, 1]\), there are constants \( c_1, c_2 \) such that when \( N \) is sufficiently large

\[
c_1 \leq \frac{\hat{a}_{Ni} - m_i}{N^{-\frac{1}{2}}(\log \log N)^{\frac{1}{2}}} \leq c_2
\]

almost surely. Therefore we come to the conclusion that

\[
\hat{a}_{Ni} - m_i \to 0
\]
at rate $N^{-\frac{1}{2}}(\log \log N)^{\frac{1}{2}}$. From (6.8),

\[ t_i - \lambda_i \to 0, \]

\[ p_i - \pi_i \to 0 \]

at the same rate. Therefore,

\[ d(\tilde{G}_N, G) \to 0 \]

at rate $N^{-\frac{1}{2}}(\log \log N)^{\frac{1}{2}}$ almost surely.

Note that under the assumptions

\begin{align*}
0 = \lambda_1 < \lambda_2 < \cdots < \lambda_{n+1} < 1; \\
0 < \lambda_1 < \lambda_2 < \cdots < \lambda_{n+1} = 1; \\
0 = \lambda_1 < \lambda_2 < \cdots < \lambda_{n+1} = 1,
\end{align*}

in a similar fashion to the situation where $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_{n+1} < 1$, we can show that with probability one

\[ d(\tilde{G}_N, G) \to 0 \]

at rate $N^{-\frac{1}{2}}(\log \log N)^{\frac{1}{2}}$. The detailed proofs for all three of these assumptions are given in Appendix 2.
Chapter 7

Future Research

7.1 Computation of MLE

There are a variety of routines for maximizing the likelihood of a mixture, such as the EM algorithm and Gateaux derivative based algorithms. One of the commonly used Gateaux derivative based algorithms is the vertex direction method (VDM). The virtues of VDM are that it increases the likelihood at each step and that it converges to the NPMLE. Unfortunately, however, the rate of convergence of VDM becomes slower and slower as the estimate approaches the maximizer of the likelihood. In addition, the number of support points of the estimate may increase unboundedly. To overcome these disadvantages of VDM we must change our point of view to seek MLE.

In Chapter 3, we explored the geometric structure associated with the maximum likelihood vector \( \hat{L} \) and gave the exact location of \( \hat{L} \) in terms of half spaces, each of which is only determined by a point on the boundary \( \partial M \). Taking this result into account, it is expected that more efficient algorithms can be developed. In addition, it is also possible to obtain other algorithms via geometrically characterizing the support hyperplane \( H \) of \( M \) at \( \hat{L} \).
7.2 Convergence Rate

We already know that the NPMLE of a mixing distribution is weakly consistent under some conditions, such as those given by Kiefer and Wolfowitz (1956), Pfanzagl (1988) and Leroux (1992). Chen (1995) established the convergence rate of the MLE in finite mixtures. To our knowledge, the convergence rate of the NPMLE in general mixtures has not yet been investigated and will be a valuable topic in the future.

For the method of moments estimator of a mixing distribution, we only gave the convergence rate for a mixture supported on $[0, 1]$ and with the number of components fixed. It is also of value to consider the convergence rate of the method of moments estimators constructed in section 5.3. This would be of aid in determining an appropriate $k_N$ therein that ensures us to solve a polynomial equation of high enough degree to achieve the desired accuracy, but not a polynomial equation of needlessly high degree.
Bibliography


    *Canad. J. Statist.* 21 139-147.

    CBMS Regional Conference Series in Probability and Statistics.

[22] Lindsay, B.G. and Basak, P. (2000) Moments determine the tail of a distribution


    Sup. Pisa Cl. Sci. (4)* 8 133-140.


[27] Morris, C.N. (1983) Natural exponential families with quadratic variance func-

[28] Pfanzagl, J. (1988) Consistency of maximum likelihood estimators for certain non-
    parametric families, in particular: mixtures. *J. Statist. Plann. Inference* 19 137-
    158.


Appendix 1

In this appendix, we prove that the limiting matrix $A$ of $A_N$ in (6.7) is invertible. If all $\lambda_1, \cdots, \lambda_{n+1} \neq 0$, then we only need to prove that the matrix

$$B(n; \lambda_1, \cdots, \lambda_{n+1}) = \begin{pmatrix}
1 & \cdots & 1 & \lambda_1 & \cdots & \lambda_{n+1} \\
2\lambda_1 & \cdots & 2\lambda_{n+1} & \lambda_1^2 & \cdots & \lambda_{n+1}^2 \\
3\lambda_1^2 & \cdots & 3\lambda_{n+1}^2 & \lambda_1^3 & \cdots & \lambda_{n+1}^3 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
(2n+2)\lambda_1^{2n+1} & \cdots & (2n+2)\lambda_{n+1}^{2n+1} & \lambda_1^{2n+2} & \cdots & \lambda_{n+1}^{2n+2}
\end{pmatrix}$$

is invertible. We will show that

$$|B(n; \lambda_1, \cdots, \lambda_{n+1})| = (-1)^{3n+5}\lambda_2^2(\lambda_2 - \lambda_1)^4 \cdots (\lambda_{n+1} - \lambda_1)^4.$$

by induction. First note that

$$|B(1; \lambda_1, \lambda_2)| = (-1)^4\lambda_1(\lambda_2 - \lambda_1)$$

and

$$|B(1; \lambda_1, \lambda_2)| = (-1)^5\lambda_1(\lambda_2 - \lambda_1)$$

for $n = 1$. For $n > 1$, we have

$$|B(n; \lambda_1, \cdots, \lambda_{n+1})| = (-1)^{3n+5}\lambda_2^2(\lambda_2 - \lambda_1)^4 \cdots (\lambda_{n+1} - \lambda_1)^4.$$
\[
(-1)^5 \lambda_1 (\lambda_2 - \lambda_1) \begin{vmatrix}
\lambda_2 - \lambda_1 & \lambda_1 & \lambda_2 \\
2\lambda_2^2 - 2\lambda_1\lambda_2 & \lambda_2^2 & \lambda_2^2 \\
3\lambda_2^3 - 3\lambda_1\lambda_2^2 & \lambda_2^3 & \lambda_2^3
\end{vmatrix}
\]

\[
(-1)^5 \lambda_1 (\lambda_2 - \lambda_1)^2 \begin{vmatrix}
1 & \lambda_1 & \lambda_2 \\
2\lambda_2 - \lambda_1 & 0 & \lambda_2^2 - \lambda_1\lambda_2 \\
3\lambda_2^2 - 2\lambda_1\lambda_2 & 0 & \lambda_2^3 - \lambda_1\lambda_2^2
\end{vmatrix}
\]

\[
(-1)^6 \lambda_1^2 (\lambda_2 - \lambda_1)^2 \begin{vmatrix}
2\lambda_2 - \lambda_1 & \lambda_2 & \lambda_2(\lambda_2 - \lambda_1) \\
3\lambda_2^2 - 2\lambda_1\lambda_2 & \lambda_2^2 & \lambda_2^2(\lambda_2 - \lambda_1)
\end{vmatrix}
\]

\[
(-1)^6 \lambda_1^2 (\lambda_2 - \lambda_1)^3 \begin{vmatrix}
2\lambda_2 - \lambda_1 & \lambda_2 \\
3\lambda_2^2 - 2\lambda_1\lambda_2 & \lambda_2^2
\end{vmatrix}
\]

\[
(-1)^6 \lambda_1^2 (\lambda_2 - \lambda_1)^3 \begin{vmatrix}
\lambda_2 - \lambda_1 & \lambda_2 \\
2\lambda_2^2 - 2\lambda_1\lambda_2 & \lambda_2^2
\end{vmatrix}
\]

\[
(-1)^6 \lambda_1^2 (\lambda_2 - \lambda_1)^4 \begin{vmatrix}
1 & \lambda_2 \\
2\lambda_2 & \lambda_2^2
\end{vmatrix}
\]

\[
(-1)^6 \lambda_1^2 (\lambda_2 - \lambda_1)^4 (-1)^1 \lambda_2^2.
\]

If we assume

\[
|B(n - 1; \lambda_1, \cdots, \lambda_n)| = (-1)^{3(n-1)+5} \lambda_1^2 (\lambda_2 - \lambda_1)^4 \cdots (\lambda_n - \lambda_1)^4.
\]

\[
(-1)^{3(n-2)+5} \lambda_2^2 (\lambda_3 - \lambda_2)^4 \cdots (\lambda_n - \lambda_2)^4 \cdots
\]

\[
(-1)^{3+5} \lambda_{n-1}^2 (\lambda_n - \lambda_{n-1})^4 (-1)^1 \lambda_n^2,
\]

then
\[
|B(n; \lambda_1, \ldots, \lambda_{n+1})| = \begin{vmatrix}
1 & 1 & \cdots & 1 \\
\lambda_1 & 2\lambda_2 - \lambda_1 & \cdots & \cdots \\
\lambda_1^2 & 3\lambda_2^2 - 2\lambda_1 \lambda_2 & \cdots & \cdots \\
\vdots & \vdots & \ddots & \vdots \\
\lambda_1^{2n+1} & (2n + 2)\lambda_2^{2n+1} - (2n + 1)\lambda_1 \lambda_2^{2n} & \cdots & \cdots \\
\end{vmatrix}
\]

\[
= (-1)^{n+3} \lambda_1 (\lambda_2 - \lambda_1) \cdots (\lambda_{n+1} - \lambda_1) (-1)^n
\]

\[
= (-1)^{2n+3} \lambda_1 (\lambda_2 - \lambda_1) \cdots (\lambda_{n+1} - \lambda_1)
\]

\[
= (-1)^{n+3} \lambda_1 (\lambda_2 - \lambda_1) \cdots (\lambda_{n+1} - \lambda_1) (-1)^n
\]

\[
= (-1)^{2n+3} \lambda_1 (\lambda_2 - \lambda_1) \cdots (\lambda_{n+1} - \lambda_1)
\]

\[
= (-1)^{n+3} \lambda_1 (\lambda_2 - \lambda_1) \cdots (\lambda_{n+1} - \lambda_1)
\]
\[
\begin{align*}
&= (-1)^{2n+3} \lambda_1 (\lambda_2 - \lambda_1)^2 \cdots (\lambda_{n+1} - \lambda_1)^2 \\
&= \begin{vmatrix}
1 & \lambda_1 & \lambda_2 & \cdots & \lambda_{n+1} \\
2\lambda_{n+1} - \lambda_1 & 0 & \lambda_2 - \lambda_1 & \cdots & \lambda_{n+1}^2 - \lambda_1 \lambda_{n+1} \\
0 & \lambda_2 & \cdots & \lambda_{n+1} \\
(2n+1)\lambda_{n+1}^2 - 2n\lambda_1 \lambda_{n+1}^{2n-1} & \lambda_{n+1}^2 - \lambda_1 \lambda_{n+1}^2 & \cdots & \lambda_{n+1}^{2n} - \lambda_1 \lambda_{n+1}^{2n-1} \\
\end{vmatrix}

&= \begin{vmatrix}
1 & \lambda_1 & \lambda_2 & \cdots & \lambda_{n+1} \\
2\lambda_{n+1} - \lambda_1 & 0 & \lambda_2 - \lambda_1 & \cdots & \lambda_{n+1}^2 - \lambda_1 \lambda_{n+1} \\
0 & \lambda_2 & \cdots & \lambda_{n+1} \\
(2n+1)\lambda_{n+1}^2 - 2n\lambda_1 \lambda_{n+1}^{2n-1} & \lambda_{n+1}^2 - \lambda_1 \lambda_{n+1}^2 & \cdots & \lambda_{n+1}^{2n} - \lambda_1 \lambda_{n+1}^{2n-1} \\
\end{vmatrix}

&= (-1)^{3n+5} \lambda_1^2 (\lambda_2 - \lambda_1)^3 \cdots (\lambda_{n+1} - \lambda_1)^3 \\
&= \begin{vmatrix}
\lambda_{n+1} - \lambda_1 & \lambda_2 & \cdots & \lambda_{n+1} \\
2n\lambda_{n+1}^2 - 2n\lambda_1 \lambda_{n+1}^{2n-1} & \lambda_{n+1}^2 - \lambda_1 \lambda_{n+1}^2 & \cdots & \lambda_{n+1}^{2n} - \lambda_1 \lambda_{n+1}^{2n-1} \\
\end{vmatrix}

&= \begin{vmatrix}
1 & \lambda_2 & \cdots & \lambda_{n+1} \\
2n\lambda_{n+1}^2 - 2n\lambda_1 \lambda_{n+1}^{2n-1} & \lambda_{n+1}^2 & \cdots & \lambda_{n+1}^{2n} \\
\end{vmatrix}

&= (-1)^{3n+5} \lambda_1^2 (\lambda_2 - \lambda_1)^4 \cdots (\lambda_{n+1} - \lambda_1)^4 |B(n - 1; \lambda_2, \cdots, \lambda_{n+1})|.
\end{align*}
\]

Thus, by the assumption, the above expression is

\[
(-1)^{3n+5} \lambda_1^2 (\lambda_2 - \lambda_1)^4 \cdots (\lambda_{n+1} - \lambda_1)^4 (-1)^{3(n-1)+5} \lambda_2^2 (\lambda_3 - \lambda_2)^4 \cdots (\lambda_{n+1} - \lambda_2)^4 \\
(-1)^{3+5} \lambda_n^2 (\lambda_{n+1} - \lambda_n)^4 (-1)^1 \lambda_{n+1}^2.
\]
Since

\[ 0 < \lambda_1 < \lambda_2 < \cdots < \lambda_{n+1} < 1, \]

\[ |B(n; \lambda_1, \cdots, \lambda_{n+1})| \neq 0, \] that is, \( B(n; \lambda_1, \cdots, \lambda_{n+1}) \) is invertible.
**Appendix 2**

In this appendix, we demonstrate that for the various assumptions on $\lambda_i$ given in (6.9), (6.10) and (6.11), with probability one, $d(\tilde{G}_N, G) \to 0$ at rate $N^{-\frac{1}{2}}(\log \log N)^{\frac{1}{2}}$.

To begin, assume that

$$0 = \lambda_1 < \lambda_2 < \cdots < \lambda_{n+1} < 1.$$ 

In this situation, when $N$ is sufficiently large, only case (i) or case (ii) can occur.

If case (i) occurs, then

$$|A| = \begin{vmatrix}
0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 \\
\pi_1 & \pi_2 & \cdots & \pi_{n+1} & 0 & \lambda_2 & \cdots & \lambda_{n+1} \\
0 & 2\pi_2 \lambda_2 & \cdots & 2\pi_{n+1} \lambda_{n+1} & 0 & \lambda_2^2 & \cdots & \lambda_{n+1}^2 \\
0 & 3\pi_2 \lambda_2^2 & \cdots & 3\pi_{n+1} \lambda_{n+1}^2 & 0 & \lambda_2^3 & \cdots & \lambda_{n+1}^3 \\
\vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\
0 & (2n+1)\pi_2 \lambda_2^{2n} & \cdots & (2n+1)\pi_{n+1} \lambda_{n+1}^{2n} & 0 & \lambda_2^{2n+1} & \cdots & \lambda_{n+1}^{2n+1}
\end{vmatrix}$$

$$= (-1)^{n+3} \pi_1 \cdots \pi_{n+1} \lambda_2 \cdots \lambda_{n+1}^2$$

$$= (-1)^{n+3} \pi_1 \cdots \pi_{n+1} \lambda_2 \cdots \lambda_{n+1}^2$$

$$\neq 0 \text{ (see Appendix 1).}$$

If case (ii) occurs, then $t_1 = 0$. The moment equations become

$$\left(\begin{array}{c}
\hat{a}_{N0} - m_0 \\
\hat{a}_{N1} - m_1 \\
\vdots \\
\hat{a}_{N2n} - m_{2n}
\end{array}\right) = A_N \left(\begin{array}{c}
t_2 - \lambda_2 \\
\vdots \\
t_{n+1} - \lambda_{n+1} \\
p_1 - \pi_1 \\
\vdots \\
p_{n+1} - \pi_{n+1}
\end{array}\right),$$
where of course $A_N$ is a $(2n + 1) \times (2n + 1)$ matrix, different from the one for the assumption where $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_{n+1} < 1$, that was discussed in section 6.4.

$$|A| = \begin{vmatrix} 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 \\ \pi_2 & \pi_3 & \cdots & \pi_{n+1} & 0 & \lambda_2 & \cdots & \lambda_{n+1} \\ 2\pi_2 \lambda_2 & 2\pi_3 \lambda_3 & \cdots & 2\pi_{n+1} \lambda_{n+1} & 0 & \lambda_2^2 & \cdots & \lambda_{n+1}^2 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 2n\pi_2 \lambda_2^{2n-1} & 2n\pi_3 \lambda_3^{2n-1} & \cdots & 2n\pi_{n+1} \lambda_{n+1}^{2n-1} & 0 & \lambda_2^{2n} & \cdots & \lambda_{n+1}^{2n} \end{vmatrix}$$

$$= (-1)^{n+2} \pi_2 \cdots \pi_{n+1} \begin{vmatrix} 1 & \cdots & 1 & \lambda_2 & \cdots & \lambda_{n+1} \\ 2\lambda_2 & \cdots & 2\lambda_{n+1} & \lambda_2^2 & \cdots & \lambda_{n+1}^2 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 2n\lambda_2^{2n-1} & \cdots & 2n\lambda_{n+1}^{2n-1} & \lambda_2^{2n} & \cdots & \lambda_{n+1}^{2n} \end{vmatrix} \neq 0 \text{ (see Appendix 1).}$$

Thus, in either case, when $N$ is large enough, $A_N$ is invertible. For the same reason as under the assumption $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_{n+1} < 1$, we have

$$d(\tilde{G}_N, G) \rightarrow 0$$

almost surely at rate $N^{-\frac{3}{2}}(\log \log N)^{\frac{3}{2}}$.

Next, assume that

$$0 < \lambda_1 < \cdots < \lambda_{n+1} = 1.$$  

For this situation, when $N$ is sufficiently large, only case (i) or case (iii) can occur.

If case (i) occurs, then

$$|A| = \begin{vmatrix} 0 & \cdots & 0 & 1 & \cdots & 1 & 1 \\ \pi_1 & \cdots & \pi_n & \pi_{n+1} & \lambda_1 & \cdots & \lambda_n \\ 2\pi_1 \lambda_1 & \cdots & 2\pi_n \lambda_n & 2\pi_{n+1} \lambda_{n+1} & \lambda_1^2 & \cdots & \lambda_n^2 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ (2n + 1)\pi_1 \lambda_1^{2n} & \cdots & (2n + 1)\pi_n \lambda_n^{2n} & (2n + 1)\pi_{n+1} \lambda_{n+1}^{2n+1} & \lambda_1^{2n+1} & \cdots & \lambda_n^{2n+1} \end{vmatrix} \neq 0.$$
If case (iii) occurs, then $t_{n+1} = 1$. The moment equations become

$$
\begin{pmatrix}
\hat{a}_{N0} - m_0 \\
\hat{a}_{N1} - m_1 \\
\vdots \\
\hat{a}_{N2n} - m_{2n}
\end{pmatrix}
= A_N
\begin{pmatrix}
t_1 - \lambda_1 \\
\vdots \\
t_n - \lambda_n \\
p_1 - \pi_1 \\
\vdots \\
p_{n+1} - \pi_{n+1}
\end{pmatrix}
$$

$$|A| = \begin{vmatrix}
0 & \cdots & 0 & 1 & \cdots & 1 & 1 \\
\pi_1 & \cdots & \pi_n & \lambda_1 & \cdots & \lambda_n & 1 \\
2\pi_1\lambda_1 & \cdots & 2\pi_n\lambda_n & \lambda_1^2 & \cdots & \lambda_n^2 & 1 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
2n\pi_1\lambda_1^{2n-1} & \cdots & 2n\pi_n\lambda_n^{2n-1} & \lambda_1^{2n} & \cdots & \lambda_n^{2n} & 1
\end{vmatrix}
$$

$$= \pi_1 \cdots \pi_n \\
1 \\
2\lambda_1 - 1 \\
\vdots \\
2n\lambda_1^{2n-1} - (2n - 1)\lambda_1^{2n-2} \\
\vdots \\
\lambda_1^{2n} - \lambda_1^{2n-1} \cdots \lambda_1^{2n} - \lambda_n^{2n-1} \cdots \lambda_n^{2n} - \lambda_n^{2n-1}
$$

$$= \pi_1 \cdots \pi_n (\lambda_1 - 1) \cdots (\lambda_n - 1) \\
1 \\
2\lambda_1 - 1 \\
\vdots \\
2n\lambda_1^{2n-1} - (2n - 1)\lambda_1^{2n-2} \\
\vdots \\
\lambda_1^{2n} - \lambda_1^{2n-1} \cdots \lambda_1^{2n} - \lambda_n^{2n-1} \cdots \lambda_n^{2n} - \lambda_n^{2n-1}$$
\[\begin{vmatrix}
\pi_1 \cdots \pi_n (\lambda_1 - 1) \cdots (\lambda_n - 1) & \lambda_1 - 1 & \cdots \\
\lambda_n - 1 & \lambda_1 & \cdots \\
\vdots & \vdots & \ddots \\
(2n - 1)\lambda_1^{2n-1} - (2n - 1)\lambda_1^{2n-2} & \lambda_1^{2n-1} & \cdots \\
\end{vmatrix} = \begin{vmatrix}
0 & 1 & \cdots & 1 \\
1 & \lambda_1 & \cdots & \lambda_n \\
\vdots & \vdots & \ddots & \vdots \\
(2n - 1)\lambda_1^{2n-2} & \lambda_1^{2n-1} & \cdots & \lambda_n^{2n-1} \\
\end{vmatrix} \neq 0.
\]

Hence, in either case, when \( N \) is large enough, \( A_N \) is invertible, which leads to a convergence rate \( N^{-\frac{1}{2}}(\log \log N)^{\frac{1}{2}} \) of \( d(G_N, G) \to 0 \) with probability one.

Finally, consider the assumption

\[0 = \lambda_1 < \cdots < \lambda_{n+1} = 1.\]

In this situation, when \( N \) is sufficiently large, any of the cases (i) through (iv) can occur.
If case (i) occurs, then

$$\begin{vmatrix}
0 & 0 & \cdots & 0 & 0 \\
\pi_1 & \pi_2 & \cdots & \pi_n & \pi_{n+1} \\
0 & 2\pi_2\lambda_2 & \cdots & 2\pi_n\lambda_n & 2\pi_{n+1} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & (2n + 1)\pi_2\lambda^{2n} & \cdots & (2n + 1)\pi_n\lambda^{2n} & (2n + 1)\pi_{n+1}
\end{vmatrix}$$


$$= (-1)^{n+3}\pi_1 \cdots \pi_{n+1} \begin{vmatrix}
2\lambda_2 & \cdots & 2\lambda_n & 2 & \lambda^2_2 & \cdots & \lambda^2_n & 1 \\
3\lambda^2_2 & \cdots & 3\lambda^2_n & 3 & \lambda^3_2 & \cdots & \lambda^3_n & 1 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\
(2n + 1)\lambda^{2n}_2 & \cdots & (2n + 1)\lambda^{2n}_n & (2n + 1) & \lambda^{2n+1}_2 & \cdots & \lambda^{2n+1}_n & 1
\end{vmatrix}$$

$$= (-1)^{n+3}\pi_1 \cdots \pi_{n+1} \lambda_2 \cdots \lambda^{2n}_n \begin{vmatrix}
1 & \cdots & 1 & 1 & \lambda_2 & \cdots & \lambda_n & 1 \\
2\lambda_2 & \cdots & 2\lambda_n & 2 & \lambda^2_2 & \cdots & \lambda^2_n & 1 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\
2n\lambda^{2n-1}_2 & \cdots & 2n\lambda^{2n-1}_n & 2n & \lambda^{2n}_2 & \cdots & \lambda^{2n}_n & 1
\end{vmatrix}$$

$$\neq 0.$$

If case (ii) occurs, then $$t_1 = 0$$. The moment equations become

$$\begin{pmatrix}
\hat{a}_{N0} - m_0 \\
\hat{a}_{N1} - m_1 \\
\vdots \\
\hat{a}_{N2n} - m_{2n}
\end{pmatrix} = A_N \begin{pmatrix}
t_2 - \lambda_2 \\
t_3 - \lambda_3 \\
\vdots \\
t_{n+1} - \lambda_{n+1}
\end{pmatrix}.$$

Thus, $$|A| \neq 0$$ using the results obtained under the assumption where

$$0 = \lambda_1 < \cdots < \lambda_{n+1} < 1.$$
If case (iii) occurs, then \( t_{n+1} = 1 \). The moment equations become

\[
\begin{pmatrix}
\hat{a}_{N0} - m_0 \\
\hat{a}_{N1} - m_1 \\
\vdots \\
\hat{a}_{N2n} - m_{2n}
\end{pmatrix}
= A_N
\begin{pmatrix}
t_1 - \lambda_1 \\
\vdots \\
t_n - \lambda_n \\
p_1 - \pi_1 \\
\vdots \\
p_{n+1} - \pi_{n+1}
\end{pmatrix}
\]

Using the results obtained under the assumption where \( 0 < \lambda_1 < \cdots < \lambda_{n+1} = 1 \),

\[
|A| = (-1)^{n+2} \pi_1 \cdots \pi_n (\lambda_1 - 1)^2 \cdots (\lambda_n - 1)^2
\]

\[
\begin{align*}
\lambda_2 & \quad \cdots & \lambda_n \\
\lambda_3 & \quad \cdots & \lambda_n \\
\vdots & & \vdots \\
\lambda_{2n-1} & \quad \cdots & \lambda_{2n-1}
\end{align*}
\]

\[
= (-1)^{n+2} \pi_1 \cdots \pi_n (\lambda_1 - 1)^2 \cdots (\lambda_n - 1)^2 \lambda_2 \cdots \lambda_n
\]

\[
\begin{align*}
\lambda_2 & \quad \cdots & \lambda_n \\
\lambda_2 & \quad \cdots & \lambda_n \\
\vdots & & \vdots \\
\lambda_{2n-2} & \quad \cdots & \lambda_{2n-2}
\end{align*}
\]

\[
= (-1)^{n+2} \pi_1 \cdots \pi_n (\lambda_1 - 1)^2 \cdots (\lambda_n - 1)^2 \lambda_2^2 \cdots \lambda_n^2
\]

\[
\begin{align*}
\lambda_2 & \quad \cdots & \lambda_n \\
\lambda_2 & \quad \cdots & \lambda_n \\
\vdots & & \vdots \\
\lambda_{2n-2} & \quad \cdots & \lambda_{2n-2}
\end{align*}
\]

\[
\neq 0 \text{ (see Appendix 1).}
\]
Finally, if case (iv) occurs, then \( t_1 = 0, \ t_{n+1} = 1 \). The moment equations become

\[
\begin{pmatrix}
\hat{a}_{N0} - m_0 \\
\hat{a}_{N1} - m_1 \\
\vdots \\
\vdots \\
\hat{a}_{N2n-1} - m_{2n-1}
\end{pmatrix} = A_N
\begin{pmatrix}
t_2 - \lambda_2 \\
\vdots \\
t_n - \lambda_n \\
p_1 - \pi_1 \\
p_{n+1} - \pi_{n+1}
\end{pmatrix},
\]

where

\[
|A| = (-1)^{n+1} \pi_2 \cdots \pi_n
\]

\[
= (-1)^{n+1} \pi_2 \cdots \pi_n \lambda_2 \cdots \lambda_n
\]

\[
= (-1)^{n+1} \pi_2 \cdots \pi_n \lambda_2 \cdots \lambda_n
\]

and

\[
1
\]

\[
2\lambda_2 - 1
\]

\[
(2n - 1)\lambda_2^{2n-2} - (2n - 2)\lambda_2^{2n-3}
\]

\[
1
\]

\[
2\lambda_n - 1
\]

\[
\lambda_2 - 1
\]

\[
\lambda_n - 1
\]

\[
0
\]

\[
(2n - 1)\lambda_n^{2n-2} - (2n - 2)\lambda_n^{2n-3}
\]

\[
\lambda_2^{2n-2} - \lambda_2^{2n-3}
\]

\[
\lambda_n^{2n-2} - \lambda_n^{2n-3}
\]

\[
0
\]
\[( -1)^{n+1} \pi_2 \cdots \pi_n \lambda_2 \cdots \lambda_n (\lambda_2 - 1) \cdots (\lambda_n - 1) \]

\[
\begin{array}{cccc}
2\lambda_2 - 1 & 3\lambda_2^2 - 2\lambda_2 & \cdots & (2n - 1)\lambda_2^{2n-2} - (2n - 2)\lambda_2^{2n-3} \\
3\lambda_2^2 - 2\lambda_n & \lambda_2 & \cdots & \lambda_n \\
\vdots & \vdots & \ddots & \vdots \\
(2n - 1)\lambda_n^{2n-2} - (2n - 2)\lambda_n^{2n-3} & \lambda_2^{2n-3} & \cdots & \lambda_n^{2n-3} \\
\end{array}
\]

\[
\begin{array}{cccc}
\lambda_2 - 1 & 2\lambda_2 (\lambda_2 - 1) & \cdots & (2n - 2)\lambda_2^{2n-3} (\lambda_2 - 1) \\
\lambda_2 - 1 & \lambda_2 & \cdots & \lambda_n \\
\vdots & \vdots & \ddots & \vdots \\
(2n - 2)\lambda_n^{2n-3} (\lambda_n - 1) & \lambda_2^{2n-3} & \cdots & \lambda_n^{2n-3} \\
\end{array}
\]

\[
\begin{array}{cccc}
\lambda_2 - 1 & 2\lambda_2 & \cdots & (2n - 2)\lambda_2^{2n-3} \\
\lambda_2 - 1 & \lambda_2 & \cdots & \lambda_n \\
\vdots & \vdots & \ddots & \vdots \\
(2n - 2)\lambda_n^{2n-3} & \lambda_2^{2n-3} & \cdots & \lambda_n^{2n-3} \\
\end{array}
\]

\[
\begin{array}{cccc}
0 & \lambda_2 & \cdots & (2n - 3)\lambda_2^{2n-3} \\
0 & \lambda_2 & \cdots & \lambda_n \\
\vdots & \vdots & \ddots & \vdots \\
(2n - 3)\lambda_n^{2n-3} & \lambda_2^{2n-3} & \cdots & \lambda_n^{2n-3} \\
\end{array}
\]
\[
= (-1)^{n+1} \pi_2 \cdots \pi_n \lambda^2_2 \cdots \lambda^2_n (\lambda_2 - 1)^2 \cdots (\lambda_n - 1)^2
\]

\[
\begin{vmatrix}
... & 0 & 1 & \cdots & 1 \\
... & 1 & \lambda_2 & \cdots & \lambda_n \\
... & 2\lambda_n & \lambda^2_2 & \cdots & \lambda^2_n \\
& \vdots & \vdots & \ddots & \vdots \\
... & (2n - 3)\lambda^2_n & \lambda^2_{2n-4} & \cdots & \lambda^2_{2n-3}
\end{vmatrix}
\neq 0.
\]

Thus, for every case, when \( N \) is sufficiently large, \( A_N \) is invertible. Therefore,

\[d(\tilde{G}_N, G) \rightarrow 0\]

almost surely at rate \( N^{-\frac{1}{2}}(\log \log N)^{\frac{1}{2}}\).