Kwame Nkrumah University of Science and Technology

Fitting Finite Mixture Model (FMM) to Frequency Data

By

Richard Anane Adortse

(BSc. Mathematics and Computer Science)

A THESIS SUBMITTED TO THE DEPARTMENT OF MATHEMATICS, KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MSC INDUSTRIAL MATHEMATICS

October 19, 2016

Declaration

I hereby declare that this submission is my own work towards the award of the MSc. degree and that, to the best of my knowledge, it contains no material previously published by another person nor material which had been accepted for the award of any other degree of the university, except where due acknowledgement had been made in the text.

Richard Anane Adortse	.....	.....
Student	Signature	Date

Certified by:

Dr. G. A. Okyere	.....	.....
Supervisor	Signature	Date

Certified by:

Dr. Richard K. Avuglah	.....	.....
Head of Department	Signature	Date

Dedication

I dedicate this thesis to my mother Madam Betty Doe, my brother Mr. Peter Adortse, my wife Mrs. Angela Adortse and my good friends Mr. and Mrs. Anati all who in diverse ways have contributed meaningfully to the completion of this work.

Acknowledgements

I am highly indepted to my mentor and supervisor, Dr. Gabriel Asare Okyere whose motivation, patience, critical questioning and assessment, and general high standard and passion for academic work has enable me to complete this work. A research of this scope could not have been completed without his guidance and motivation. I am really honoured to study under you.

A special note of appreciation is also extended to all lecturers in the Department of Mathematics at KNUST, particularly to Prof. S. K. Amponsah for his insightful guidance and suggestions. A warmly thanks goes to my colleagues and group members especially to Mr. Francis N. O.Gyimah.

I want to acknowledge Prof. Peter D. M., (Department of Mathematics, Mcmaster University, Canada) who constantly guides in writting the R codes and resolving critical challenges encountered when using his misdixt package. Prof. I really thank you.

I also wish to express my sincere thanks to Dr. Seth K. Agyakwah (CSIR-WRI, Akosombo) whose continuous passion in using Mathematical and Statistical models in fisheries led to this work and to the Director of the fisheries commission of Ghana, Dr. Paul Bannerman who gladly provided information and the dataset on the yellowfin tuna.

Finally, to my entire family especially my wife Mrs. Angela Adansi Adortse for her patience and understanding, my brother Peter Adortse for his support and believe in me and to my good friends Mr. and Mrs. Anati who have and continue to impart my life fantatiscally. Thank you all and God bless you.

Abstract

One of the main variables needed in analyzing quantitative growth, mortality, and stock assessment models in fishery has been age data. Unfortunately, age as the biological measure of time is not readily available in Ghana and other tropical countries. Also, the graphical based methods used by most fishery scientists to dissect and estimate demographic parameters from fishery length frequency data (a viable substitute for the age data) makes statistical inference unreliable. In this work, the computational method of FMM is used to decompose the 2014 yellowfin tuna population into components (age-groups) and the estimates of each component obtained by the Maximum Likelihood (ML) method via Expectation Maximization (EM) and Newton Raphson (NR) algorithms. Based on the size selection in this fishery, it could be infered that the yellowfin tuna population consists of five subpopulations. The mixing proportions for successive age groups increase continuously until at least the fourth component. The difference in the continuous increment in the mean-lengths is at least twice the standard deviations. For efficient decomposition and estimation of the mixture parameters, the study recommeds the FMM and ML method via EM and NR algorithms over traditional graphical methods.

Contents

List of Figures

List of Tables

List of Abbreviations

ANOVA	Analysis of Variance
CCV	Constant Coefficient of Variations
EM	Expectation Maximization
ESM	Equally Spaced Means
FMM	Finite Mixture Model
GCM	Growth Curve Means
GFMM	Gaussian Finite Mixture Model
GOF	Goodness of Fit
ML	Maximum Likelihood
NR	Newton Raphson
SE	Standard Error
SMF	Specified Means Fixed
SPF	Specified Proportion Fixed

Chapter 1

1.0 Introduction

The yellowfin tuna is one of the most important species of fishes found in the atlantic ocean of Ghana. Its contribution to job creation, nutrition, and gross domestic product (GDP) is very significant. It is the main raw material used by pioneer food canery in Tema, and also served as a source of employment to some local folks. The good news is that, fishery resources and for that matter the yellowfin species are renewable unlike mineral resources. Hence, the need for quantitative and scientific information to be made available for prudent management of this important species.

Age data and length-at age data have been considered the most precise and commonly used data in quantitative fishery assessment models such as stock, growth, and mortality rate. But in the tropics like Ghana, reading the precise age of yellowfin tuna by observing calcified structures such as hard parts (otoliths) is complex and costly because sensitive microscopes need to be used. Fortunately, due to the large availability of length frequency data on yellowfin tuna from the database of Fishery Commission of Ghana, meaningful demographic parameters such as the age-groups, mean lengths, proportions, and standard deviations can be estimated by the maximum likelihood method from Gaussian finite mixture model.

1.1 Background

In heterogeneous or mixed population, if the number of components and the underlying distributions are known, then finite mixture model (FMM) can be fitted to the sample meaningfully. But in practice, the number of components are unknown and therefore need to be estimated or assumed based on previous work in similar population. The distribution is also assumed to come from the exponential family particularly the normal distribution.

Generally, the finite mixture distribution model is define as: a compounding distribution which consists of a finite (k) components of distributions sampled from an unknown number of subpopulations (heterogeneous) with each component distribution assuming a certain probability density functions (pdf), $f_i(x)$ and mixing proportions π_i. In theory, the components pdfs $f_i(x)$ may be consisdered differents. However, in practice, the same form is frequently used, the most popular being the Gaussian or the log normal.

For the estimation of the mixture parameters from FMM to be meaningful, the necessary and sufficient condition of $identifiability$ have to be satisfied. The GFMM has been demonstrated to satisfied this critirion (Yakowitz 1969). In principle, this condition gurantees that, all the parameters in each component can be estimated from the normal finite mixture model. That is, given two distinct mixture densities, $g_1(x|\boldsymbol{\Psi}_1)$ and $g_2(x|\boldsymbol{\Psi}_2)$ being equal, the two sets of the parameter values $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$ are also equal. Contrapositively, no two sets of distinct parameter values will produce the same mixture density function $g(x|\boldsymbol{\Psi})$.

The application of this important distribution model to incomplete data problems (grouped, truncated, missing, etc.) dated back to the works in the last century. Pearson (1894), published a paper on estimating the parameters of two components Gaussian distribution. In recent times, it has been applied in several fields such as: medical, biology, etc. For instance, in this study, the GFMM is used to dissect and estimate the yellowfin tuna demographic parameters.

There are various methods for estimating the mixture parameters from incomplete data problems. These includes: the methods of moment, the maximum likelihood (ML), minimum Chi-square, and Bayesian method. Macdonald and Pitcher (1975) in their paper showed that the minimum Chi-square and the ML methods yield equivalent estimates in large sample. In this study, the ML method via two iterative procedures: EM algorithm and the NR algorithm are used.

The primary concept of the ML estimation method is to find the parameter estimate that maximizes the log likelihood function of the finite mixture model given in a given grouped data. In practice, this log likelihood function is analytically intractable (lack closed forms). This is due in part to the mixed data problem and also in part to the grouped data problem. In such cases, iterative procedures (EM and NR) are use to approximate the estimates.

The EM algorithm is a standard tool used to approximate intractable ML estimators in finite mixture models. The general concept behind this algorithm is to convert the maximization problem involving the complex incomplete log likelihood function into a complete data log likelihood whose gradient or score equation can be approximated iteratively with some appropriate initial values. Its slow convergence property and inability to authomatically compute the variance covariance matrix are among the obvious weaknesses of this algorithm. Dempster el al. (1977) claimed that the roots to the M-step of the EM algorithm applied to incomplete data cannot be solve analytically since the function from the E-step is not in a closed form.

Various methods to solve this problem have been proposed. Wengrzik et al. (2011), in their paper presented five comparative methods using simulation study to find the roots to the M-step. The Macdonald et al. (1975), approach implemented in the *mixdist* R package was considered the most outstanding method and package for modeling mixture problems with grouped data. In this method, NR algorithm was employed to solve the intractable score equation generated in the E-step of the EM algorithm by approximating the gradient function with

some Taylor series expansion. The NR algorithm converges quicker and also, authomatically estimate the error covariance matrix. However, it is considered to be very sensitve to the choice of the initial values and difficult to implement as a stand alone algorithm. This is the method used in this work.

Furthemore, a data problem is said to be incomplete data problem (finite mixture with grouped data) if the components the individual observations belong to are not observed directly but only the marginal distribution of the observations are observed. In contrast, in a complete data problem, both the individual observations and the components the distribution belong to are observed.

Finally, a finite mixture data is said to be grouped if the individual observations are not reported but rather the number of observations (frequencies) falling into the respective class intervals with fixed boundaries are reported.

1.2 Problem Statement

In order to judiciously manage the exploitation of the different tuna species in the atlantic ocean, fishery scientists and policy makers over the years have relied on growth, mortality and stock assessments modeling. The main variables used in these modelings have been age and length-at-data of the particular species. However, in the tropics like Ghana, reading the precise age of yellowfin tuna by observing calcified structures such as hard parts (otoliths) and scales are very complex and costly since sensitive microscopes need to be employed. In the absence of the individual age data, the size structure (length or weight frequency data) is considered a viable replacement commonly used in fishery assessment models (Newman, et al. 2007). Length increments as the biological measure of time in the juvenial stages of fishes are considered to be proportional to age or growth increments. Also, length frequency data are easy and cheaper to collect. Hence, large samples are readily available.

The traditional parametric method used by most biologists in analysing (dissect-

ing and estimating fishery demographic parameters) from length frequency data has been variants forms of graphical methods. See the works of Petersen (1891) and Bhattacharya (1967). These methods make meaningful statistical inference unreliable. For instance, the interpretation of the estimates from the inspection of modes (Petersen 1891) and the inflexion points (Bhattacharya 1967) from the probability plots are strictly subjective. The computational FMM method of dissecting and the ML method of estimating the mixture parameters are considered by far more statistically significant.

The study anticipates to find: the number of subpopulations in the 2014 yellowfin tuna population from the length frequency data and to estimate the demographic parameters of each component. That is : the age-groups, proportions, mean-lengths, and standard deviations by the ML estimation method via EM and NR algorithms. The MacDonald and Du (2011) *mixdist* package in R was used in the fitting. Furthermore, for meaningful inferencing and diagnosis, the Chi-square goodness of fit to test the assumption of normality and the complete covariance matrix of the estimates will be computed.

1.3 Objectives

The aim of this project is to apply the GFMM to the yellowfin tuna grouped data and also to estimate the parameters of the mixture model (age-groups, mixing proportions, mean lengths, and standard deviations) using the ML method through the EM and NR algorithms.

This study specifically seeks to:

1. find the number of subpopulations (age groups) by dissecting the 2014 length frequency data into its constituent components

2. estimate the parameters of each component by the ML method via the EM and NR algorithm.

3. obtain the complete variance covariance asymptotic matrix and the Chi-square goodness of fit test.

1.4 Methodology

Due to the mathematical and statistical complexity in dissecting and estimating mixture parameters from length frequency data, fishery biologists have over the years relied on less efficient graphical methods such as the Bhattacharya method for the decomposition and the estimation of the cohorts parameters. The problem at hand is to dissect into components and estimate all the parameters of each component using the GFMM from yellowfin tuna length frequency by ML method in conjunction with two iterative algorithms: EM and NR algorithms.

The data used for the model was 2014 monthly length frequency yellowfin tuna data from the atlantic ocean of Ghana collected from the log books of the fishery commission of Ghana. These species and the year was chosen because of their high market value and their abundance in 2014. The data was first grouped into class intervals of equal width of 2cm and the number of yellowfin falling into each interval was reported. The histogram was chosen for the exploratory data analysis.

1.5 Significance of Study

Several studies on length frequency analysis of fishery species particularly using the graphical methods like Bhattacharya have been conducted by fishery scientists in Ghana. However, there is very little or no detail statistical analysis obtained from applying FMM to dissect and estimate the mixture parameters from length frequency. This could hinder meaningful demographic information (growth, mortatlity, and stock assessment) about these important species to policy makers in the industry. All such stock assessment and other complete analysis of the tuna

species are performed outside the country, particularly in Portugal.

Now that the commission is planning to perform all these analysis within, findings from this research could fill the temporal gap in making information available to decision and policy analyst locally. The result from this findings could also be adopted in regulating fishing activities of the various vessels that docked at the Tema harbour.

The study was restricted to the atlantic ocean of Ghana. The random sample was collected from landings of various vessels that docked at the Tema fishing Harbour in 2014. It consisted of the fork length FL (in cm) of yellowfin tuna sampled monthly from the Eufra fleet using the PS gear and free school samplying method. This data was made available by the fishery commission of Ghana.

1.6 Limitation of Study

The research intended to model the individual growth rate of the yellowfin tuna using the Von Bertanllanfy growth model to the monthly length at age data. However, the individual age data is not available in Ghana. Hence the mean lengths (which could be interpreted as the mean growth rate) at successive age groups was estimated from GFMM. The correct number of components could not be estimated by our method, and hence was guessed based on the visual inspection of the histogram. The appropriate initial values for the iterative algorithms should have come from a small subsample of direct age readings which is unavailable in our case. Again, our initial values were purely guessed through trial and error aided by the sample histogram.

Finally, a lot of constraints were imposed in order to estimate all the parameters of each component meaningfully.

1.7 Thesis organization

The rest of this study is organized as follows: Chapter two contains review of related literature or abstract relating to length frequency analysis of the yellowfin tuna, finite mixture distribution model, ML estimation method, EM algorithm, and NR algorithm. In Chapter three, the theories behind the GFMM and the estimation methods of mixture parameters particularly, the ML via EM and NR are examined in detail. Furthermore, the error covariance matrix and the Chi-square goodness of fit test as diagnostic measures are discussed. The analysis and discussion of the estimates from the fitting of the GFMM to the yellowfin length frequency data is carried out in chapter four, and finally, chapter five presents the appropriate conclusion and recommedation of the study.

Chapter 2

LITERATURE REVIEW

2.0 Introduction

This chapter reviews relevant literature on the theory and application of the GFMM in various fields particularly in fisheries. In particular, the chapter reviews related papers on the various estimation methods for mixture parameters particularly estimation by ML method via Expectation Maximization and Newton Raphson algorithms.

2.1 General Background on Finite Mixture Model (FMM)

The study of fitting mixture distribution model to subpopulations sample dated back to the last century. Pearson (1894) estimated two proportions by the method of moments from a histogram of two normal mixtures. His sample size of 1000 crabs was obtained from the measure of the ratio of forehead to body length. With the popularity of high speed computers, finite mixture distribution models have become the standard statistical model for fitting problems in heterogeneous or a number of homogeneous populations in several fields: in fishery, biology, economy, etc. For instance, in fishery, the mixture model is used to dissect a

number of homogeneous or heterogeneous populations into constituent components (age groups). This is very crucial to the fishery worker since indispensible age data used in stock assessment and other related models are very difficult, costly, and time consuming to obtain directly from calcified structures especially in the tropics like Ghana (MacDonald and Pitcher, 1979; Pauly and Mogan, 1987).

The extensive background and literature on the theory and application of the FMM are reviewed by Everitt and Hand (1981), Titterington et al. (1985), MaLachlan and Basford (1988), Lindsay (1995), Bhoning (1991), and MaLachlan and Peel (2000).

In practice, the distribution of the mixture density is not known, but only assumed or chosen based on the histogram. The fittings of some of these models focus on distributions assumed to come from the discrete family of which the poisson and the binomial are the commonest (Blischke, 1965 and Schilling, 1947) , whereas the majority of the fitting assumed the Guassian or the easily transformed log normal distribution. Everitt and Hand (1981) and MaLachlan and Basford (1988) argued that the choice is purely due to popularity , convenience, and desirable statistical properties.

MaLachlan and Peel (2000) detailed historical review of the FMM since the last century revealed that the evolution of the FMM was necessitated by the difficulty in implementing Pearson's estimation method of moment in many empirical situations. This was due partly to the mathematical complexity in the mixture denstity functions and also partly to the difficulty to program this method on low speed computers.

2.2 Estimation Methods of FMM

The absence of high speed computers together with the intractability of the methods of moment paved way for various statistical computational and graphical

estimation methods (Fowlkes, 1979) all assuming a certain distribution. The graphical methods entail plottings on normal probability paper and the most popular one has been the Bhattacharya method. The computational methods consist of the methods of moment, various minimum distance methods, the ML method, and the Bayesian method.

A comprehensive review of the various graphical methods were included in the works of Buchanan-Wollaston and Hodgson (1929), Harding (1949), Preston (1953), and Cassie (1954). Related improved graphical methods can be found in the works of the following authors: Hald (1952), Oka (1954), Tanaka (1962), Taylor (1965), Bhattacharya (1967), and Harris (1968).

The graphical methods were introduced in the last century by Petersen (1891). His method consisted of the processes of graphically dissecting length sample into components or cohorts and then identifying the peaks of each relative age of the fish. Afterwards, the peaks are then linked together in a process called modal progression analysis. Pauly and Gayanilo (1997) claimed that one weakness in this method is that both the processes of identification and connection of the peaks cannot be repeated with precision. The interpretations are also purely subjective and less reliable. Thus, make statistical inference unreliable.

In recent times, the Bhattacharya (1967) method has been the most popular graphical method used in the biological and fishery sciences. This method assumed that fish lengths follow the normal or the log normal distribution before dissecting the lengths into age-groups or cohorts each representing a component of the mixture. This was achived by transforming the normal mixture distribution problem (non linear problem) into a linear, and then computing the estimates from a regression analysis. Kolding and Ubal Giordano (2002) elaborated on the steps in transforming the normal distribution to a linear before the computation of the mean lengths, proportion sizes, and the standard deviations from the regression equation. Pauly and Caddy (1985) claimed that this method is less subjective and cumbersome compared to the Pertersen method.

Pearson (1894) was the first person to apply the method of moments to obtain the five parameter estimates of a two mixture components distribution. Since then, this method of estimation has gained an extensive application in various fields. Pearson proposed comparing all the potential solutions and selecting the best fit as the solution to the multiple critical values problem of the likelihood function.

Another method of estimation of mixture parameters that has received considerable attention in practice is the various minimum distance methods. See Titterington et al. (1985) and McLachlan and Peel (2000). A version of this method that has been reviewed in most journals is the Hellinger distance minimum distance estimation method. Wordward et al. (1983, 1984, 1990) applied the Hellinger distance method to estimate the parameters (proportions) of two components normal mixture distribution. They demonstrated that the Hellinger minimum distance estimator is statistically efficient and robust compared to the Cramer-Mises minimum distance method. Beran (1977) also theoretically confirmed that the Hellinger minimum distance method is a viable alternative in certain situations.

The most popular and standard method of finding the parameter estimates in mixture models has been the ML method via EM-algorithm. It is argued that the ML method is asymptotically consistent, efficient, and unbiased (Ojeda, 2001). That is, it is statistically stable and converges optimally. However, Anandkumar et al. (2012) argued that this method is computationally complex and likely to diverge or fail when the number of components become very large even with advent of high speed computers. He demonstrated that the method of moments is a viable empirical alternative in multi-mixture and hidden Markov models situations where the number of components and hence the parameter estimates increases.

Unlike the ML estimation method which maximizes the log likelihood density function, the various minimum distance estimation methods attempt to minimize

the sum of the squares of the distance between the theoretical denstity distribution function and its empirical funtion. Quandt and Ramsey (1978) proposed the use of the moment generating functions of the theoretical and the empirical functions respectively.

Wirjanto and Xu (2009) in their survey, applied moment generating function minimum distance estimation method to stock data finance. They demonstrated that this method overcame the initial requirement of the log likelihood function in the ML method to be bounded over the parameter space. They concluded that in empirical situations, where most of the log likelihood functions are not bounded, a viable alternate approach is to apply Qandt and Ramsey (1978) moment generation function and Schmidt (1982) modified moment generating function to the traditional minimum distance method.

With the popularity of high speed computers, mixture distributions have become the defacto statistical distribution for fitting problems in heterogeneous populations in several fields: in fishery, biology, economy, etc.

In their paper, Macdonald and Pitcher (1975) asserted that the histogram should be chosen as the standard representation of size frequency data over traditional graphical methods used by most biologists. They also demonstrated that the minimum Chi-square estimators and the ML estimators are asymptotically stable and equivalent when applied to grouped data. In large grouped sample, minimizing the log likelihood density is prefered to maximizing it. They argued that, the interpretation of the discrepancy between the mixture density and the empirical denstiy is easy and meaningful, and also, the minimized estimator can easily be used in the computation of the Chi-square goodness of fit test.

The ML method has been argued to be the most popular or widely used of all estimation methods. This method was first advocated by Rao (1948). He applied the ML method through iterative procedure making use of Fisher's scoring method to approximate the parameters in two component univariate mixture distribution. Other authors like Baker (1940) and Mendenhall and Harder (1958) obtained

mixture estimates through similar iterative approach. For instance Mendenhall and Harder obtained the mixture estimates by applying the the Newtons method as the required iterative procedure.

A thorough application of the ML estimation method are captured in the works of Everitt and Hand (1981) and Titterington et al. (1985). Under certain conditions, the ML estimators have been proven to be theoretically friendly and statistically efficient, consistent, and robust compared to other methods. One major weakness in this method was that, it was computationally tedious to implement prior to the popularity of high speed computers (Everitt and Hand, 1981).

The advent of high speed computers couple with powerful iterative and numerical procedures have seen a renew interest in the application of the ML estimation method by researchers. Complex and intractable computations can now easily be programmed on computers (Brady, 2008; Redner and Walker, 1984).

Redner and Walker (1984) wrote an extensive review on the various estimation methods focusing especially on the minimum distance method and the ML methods. They commented that, of all the various iterative procedure avalaible, the EM-algorithm is by far the most popular and widely used procedure to approximate intractable mixture parameters. Less sensitivity to initial values and easy programmability on computers are some of its advantages. However, the EM-algorithm have been demonstrated to converge really slow due to its linear rate of convergence compared to NR-algorithm quadratic convergence rate.

Dong (1996) in his paper cited instances of earlier application of the ML estimation method in various fields. For instance, he commented that Do and McLachlan (1984) applied mixture analysis to owl population and hence was able to estimate the proportions and the diet of the owl.

For the decomposition of subpopulations into constituent components to be statistically meaningful, the trivial choice of distribution have been the mixture distribution. However, one major challenge faced by researchers have been which particular estimating method is more viable. Holgersson and Jorner (1978) in

their paper gave a detailed theoretical and empirical account of the all the estimation methods of mixture models.

Day (1969) examined the problem of estimating mixture parameters from the normal distribution by comparing the methods of moment, minimum Chi-square, Bayesian, and the ML methods under simulation study. He concluded that the first three methods are inferior to the likelihood method partly due to the computational complexity and also partly to weaknesses in their sampling properties except in univariate cases.

2.2.1 Expectation Miximization (EM) Algorithm

Prior to the formal introduction of the EM-algorithm by Dempster et al. (1977), Sundberg (1974, 1976) introduced the iterative method of approximating roots from ML equations encountered in most incomplete data problems. Dempster et al. (1977) was accredited with formally introducing the EM-algorithm based upon Sundberg's work.

Sundberg (1974, 1976) elaborated on the incomplete data problem as data coming from mixture population, missing data, grouped, censored or truncated data which are usually associated with exponential family of distributions. According to him, the Newton-Raphson and the Fisher's scoring methods used as the standard numerical procedures at the time in solving the intractable likelihood equations are often asymptotically instable. The initial matrices inversion of these procedures are tedious and failed in many empirical situations, and also, the starting values are sensitive and hence does not often converge. He therefore proposed an iterative procedure that is simpler, less sensitive, and converges quickly based on the works of Blight (1970) who failed to prove its convergence.

According to Dempster el at. (1977), the EM algorithm is an iterative procedure used in estimating ML parameters of data observations considered to come from incomplete data problem. In the expectation (E-step), a certain $Q(.)$ function is generated using the conditional expectation of the complete log likelihood data

given the observed with the appropriate initial values. A transformation of the incomplete log likelihood yields the required complete data likelihood. The M-step then maximizes the $Q(.)$ function generated in the E-step iteratively until convergence or a certain tolerance is reached. They argued that, the EM algorithm is simpler to implement, less sensitive to initial values, and has broad area of applications. Apart from its slow convergence, EM algorithm has the weakness of not being able to solve all incomplete data problem. It is also argued that the EM-algorithm lacks the computational power to directly estimate the covariance matrix.

In order to speed up the rate of convergence and also solve the analytically intractable equation at the M-step, variants forms of Newton's numerical procedure such as the quasi Newton, Newton Raphson method, and conjugate gradient acceleration have been proposed. Jamshidian and Jennrich (1977, 1993) applied new hybrid quasi-Newton to poisson mixtures and the conjugate gradient acceleration to the EM-algorithm to estimate the covariance matrix from incomplete multivariate normal mixtures respectively. They concluded that the new hybrid quasi-Newton method accelerate the convergence rate of the EM-algorithm in excess of over 50 factors and also the conjugate gradient acceleration method speeds up the convergence rate in over 10 factors.

Recently, Du (2002) in his thesis and R-package, R*mix* applied the combination of EM-algorithm and NR-algorithm based on MacDonald and Green (1988) mix software to northern pike fishery dataset. He concluded that NR-algorithm applied to the M-step of the EM-algorithm after appropriate initial values have been chosen in the E-step speed the convergence of the EM-algorithm tremendously and also help in approximating the analytically intractable log likelihood equation in the M-step. MacDonald and Du (2011) R-package "*mixdist*" which is used in this thesis was based on the works of the former two authors.

One of the basic requirements for the convergence of the EM-algorithm is the monotonicity (non-decreasing) property of the log likelihood denstity function.

See Dempster et al., (1977), Redner and Walker (1984), Meng and Rubin (1993), Chen and Gupta (2010), Wu (1983), Xu and Jordan (1995), etc. That is, if the $(m+1)^{th}$ parameter iterate of the $Q(.)$ function in the M-step increases continuously than the current m^{th}, then the log likelihood of the $(m+1)^{th}$ parameter is guarantee to be non-decreasing (increases) in contrast to the log likelihood of the m^{th}. Chen and Gupta (2010) argued that, the monotonic property together with appropriate initial values under some regularity conditions of the conditional statistics, $Q(.)$ in the E-step are the necessary requirements to guarantee convergence.

Wu (1983) answered two critical questions ignored by Dempster et al. That is, whether the log likelihood of the parameter converges to global maximum, local maximum or some saddle points. He demonstrated that general convergence to global maximum is not feasible. Convergence to local maximum is only possible under mild conditions, but he failed to demonstrate it empirically. However, together with some other conditions, he proved that the log likelihood converges to some critical or stationary values.

Chen and Gupta (2010) in their tutorial, give a detail mathematical background of the EM-algorithm. Applications of the GFMM and Hidden Markov model (HMM) were carefully discussed. To overcome the weaknesses and near to failure (divergence) of the EM-algorithm due to non-concavity and singularities of the log likelihood function resulting from lack of boundedness, Chen and Gupta (2010) proposed the use of global optimizers like the Newton Raphson algorithm in conjuction with the EM-algorithm. Also, the Bayesian estimation method could be employed to overcome the problem of singularities of the log likelihood.

Couvreur and Bresler (1995) applied the ML method via EM-algorithm to decompose a mixture consisting of finite length and authoregressive model in order to determine and classify the number of authoregressive signals of unknown variances. They used the methods of moment to obtained the initial values for the EM-algorithm. The comparative advantages of applying the EM-algorithm to

incomplete data problem to other numerical methods like NR and Scoring methods has been studied by Couvreur (1996). He argued that, the initial values of the numerical methods are analytically complex to guess and also exhibit convergence instabilities when the number of parameters to be estimated are high. He demonstrated the application of the EM-algorithm to estimate the parameters of a poisson mixture of positions emission tomography. Several acceleration schemes have been derived to address the slow convergence property of the EM-algorithm and also to approximate the indirect covariance matrix.

One of the unresolved challenges in applying the ML method in mixture modeling situations have been how to estimate the correct number of components or classes (subpopulations). See Nylund et al. (2001). In their simulation study, they examined and compared the likelihood based tests, Information Criterion like the Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC) used by many researchers. They concluded that, the Boostraped Likelihood Ration Test (BLRT) applied to different mixture problems is superior followed by the Bayesian Information Criterion (BIC), and then the adjusted BIC.

Hathaway (1983, 1985) in his Phd thesis and article attempted to solve the problem of divergence usually encountered in empirical situations when the ML method is used to estimate mixture parameters. He proposed applying basic constraints to the parameter set (standard deviations, mean-lengths and the proportions) will transform the mixture problem into a well poised optimization one. He concluded that the EM-algorithm when applied to a constraint mixture problem will yield consistent and efficient estimates.

MacDonald and Pitcher (1979) imposed the following basic constraints: the mixing proportions or weights to sum to one and to fall between zero and one, the standard deviations to be positive and the consecutives means to be strictly increasing. They argued that these basic constraints together with the knowledge of the distribution of each k subpopulation in the mixture model, will ensure that the necessary condition of identifiability of the GFMM to be satisfy as showed

by Yakowitz (1969). MacDonald and Green (1988) and Du (2002) expanded on these basic constraints to include combinations of several other constraints on the proportions, standard deviations, and the means all being dictated by the particular problem encountered by the researcher. Without appropriate constraints on a particular problem, the estimation of all the parameters will not be meaningful especially when the components are heavily overlapped and large.

Most datasets used in both supervised and unsupervised learning come from an incomplete data source partly due to missing data values. To be able to learn from these incomplete datasets, FMM are considered the standard model. Likewise, the ML estimation method via the EM-algorithm has been the standard method used to estimate the mixture parameters. (Ghahramani and Jordan, 1994).

Laslett et al. (2004) estimated the GFMM parameters (mean lengths and standard deviations) from the South Australian bluefin tuna length frequency data by the ML method. The estimates were obtained through a first and second derivative optimization method implimented by Polacheck et al. (2003). They argued that the decomposition is only meaningful in juvenial bluefin tunas together with the assumption that spawning activities is within a limited biological time. In older fisheries, there is a heavy overlapping of the components and hence estimation could be problematic.

Similarly, Zhu and Zhao (2013) applied MacDonald (1987, 2008) and Du (2002) Mix and Rmix techniques (packages) to decompose the American Eel into components (cohorts) and was able to estimate the mean lengths and the age groups of the Eel.

Haitovsky (1983) in his paper gave a concise definition of a grouped data and summarized the major reasons for grouping data. He cited among other reasons: for easy descriptive purposes as in the graphical or tabular displays and data privicy requirements in analysing confidencial data. Also, for easy computation and storage in situations where the datasets are really large. Lastly, due to the inaccuracies in measurements that could result from imprecision of some

instruments, subjectivity of the researcher in taking the readings, etc.

One major challenge in implementing the grouped data densities has been the disparities in the various estimation methods, and hence the inferences of the parameters. Heijan (1989) compiled a comprehensive historical literature on the various estimation methods. He recommended the ML based methods and the recent Bayesian method to the method of moments.

Some obvious weaknesses encountered when applying the EM-algorithm to incomplete data have been : it has a relatively slow convergence rate (linear rate), and the inability to directly compute the standard error or the covariance matrix. Louis (1982) attempted to overcome the problems encountered in the EM-algorithm based on earlier historical reviews by proposing procedures that speed the convergence rate and also estimate the information matrix. He solved the information matrix problem by applying Woodbury (1977) missing information principle to the computation of the gradient vector and the negative Hessian matrix of the associated complete-data.

For a successful implementation of the EM-algorithm to estiamte the mixture parameters, the number of components (age-groups) have to be known in advance. However, in real situations, only the data through the histogram is available which is not reliable. The components from the histogram are often overlapped and do not all form distinct modes. MacDonald and Pitcher (1975) proposed direct ageing of a small sub-sample in fisheries to estimate the number of age-groups and some initial parameters of the mixture model. On the other hand, Wang et al. (2004) proposed a computational procedure, stepwise split-and-merge EM (SSMEM) algorithm based on two posterior probability critiria that could simulteneously estimate the number of components and the parameters of the mixture model. Their implementation revealed that the SSMEM is statistically robust and significant.

Chapter 3

METHODOLOGY

3.0 Introduction

This chapter discusses the mathematical and the statistical theories behind the GFMM and the various estimation methods of estimating the mixture parameters. In particular, the ML method from incomplete data problems is examined in detail. The iterative procedures: EM and NR algorithms used to approximate the analytically intractable log likelihood function will also be discussed briefly. Finally, the complete covariance matrix of the mixture problem applied to grouped data will be deduced.

3.1 Finite Mixture Distribution Model

A finite mixture distribution model is a standard and a very flexible model used in modeling data that are encountered in many practical situations such as: data from heterogeneous population, complete and incomplete data problems (missing data problems, truncated, and censored data) situations. Its worth to statiscians and researchers is underscored by the volume of articles published on it.

Generally, it is define as a compounding distribution which consists of a finite k subpopulations (components) of distributions sampled from a number of homoge-

neous or heterogeneous populations with each component distribution assuming the same or different density function, $f_i(x)$. This important model is applied in various fields such as: biology, astronomy, medicine, engineering, marketing, economics, etc. For instance, in biology it is used in the determination of sex and age groups parameters from size frequency data and in the medical field, it is used in the diagnosis and prognosis in patients.

Definition: 3.1

Let X be a k component random variables in the sample space $\mathcal{X}$. Furthermore, supposed $f(x|\boldsymbol{\theta}_i)_{,\ i=1,\ldots,k}$ denote the mixture components densities with parameter vector $\boldsymbol{\theta}_i$ consisting of the parameters (μ_i, σ_i), then the finite mixture density function is defined as :

$$\begin{aligned} g(x|\boldsymbol{\Psi}) = &\ \pi_1 f_1(x) + \cdots + \pi_k f_k(x) \\ = &\ \sum_{i=1}^{k} \pi_i f(x|\boldsymbol{\theta}_i) \qquad (x \in \mathcal{X}), \end{aligned} \tag{3.1.1}$$

where π_i are the mixing proportions or weights and are constrained to $0 \leq \pi_i \leq 1$,$(i = 1, \ldots, k)$ with $\pi_1 + \cdots + \pi_k = 1$ and $\boldsymbol{\Psi} = (\pi_i, \ldots, \pi_k, \boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_k^T)^T$ is the complete collection of all distinct parameters occuring in the mixture model.

Assuming that all the $f_i(x|\boldsymbol{\theta}_i)$ are normally distributed, but with different means and standard deviations, then $f_i(x|\boldsymbol{\theta}_i)$ is define as:

Definition 3.2

Let X be a random variable with a parameter vector $\boldsymbol{\theta} \in (\mu_i, \sigma_i)$, then the Gaussian or normal distribution density function of X is given by

$$f(x|\boldsymbol{\theta}_i) = f(x|\mu_i, \sigma_i) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}[(x - \mu_i)/\sigma_i]^2 \right\} \tag{3.1.2}$$

where μ_i and σ_i denote the mean and standard deviations of the $i'th$ component

distribution.

3.2 Other Constraints and Conditions Necessary for a meaningful estimation of mixture parameters

In addititon to the basic constraints on the mixing proportions in definition 3.1, Macdonald and Green (1988) imposed variants forms of constraints on the proportions, means, and the standard deviations. Also, the necessary condition for the estimation to be meaningful is that the mixture densisty function must be *identifiable.*

3.2.1 Constraints on the Proportions

In empirical situations, the estimation of all the mixture parameters are very difficult. Hence it becomes necessary to impose some basic constraints in order to meaningfully estimate all the parameters.

3.2.1.1 Proportions Free (None)

Here, only the basic constraint $0 \leq \pi_i \leq 1$ $(i = 1, \ldots, k)$ with $\pi_1 + \cdots + \pi_k = 1$ is imposed. The first $(k-1)$ proportions will be estimated directly with the π_k^{th} proportion calculated from the relation

$$\pi_k = 1 - \sum_{i=1}^{k-1} \pi_i \tag{3.2.1}$$

3.2.1.2 Specified Proportion Fixed (SPF) Constraint

Here, some or all of the proportions are held fixed in addition to the basic constraint while the other parameters are estimated. Suppose n is the number of

proportions held fixed, then the number of free proportions to be estimated is $k-(n+1)$. All the proportions can be made equal by evaluating each one at $1/k$.

3.2.2 Constraints on the Means

3.2.2.1 Means Free (None)

The constraint

$$\mu_1 < \mu_2 < \ldots < \mu_k \tag{3.2.2}$$

is imposed in order to uniquely estimate all the means directly. This is to avoid multiciplity of the mean estimates. This is the constraint that is employed in the current work.

3.2.2.2 Specified Means Fixed (SMF)

Some means are held at fixed values while the remaining are estimated

3.2.2.3 Equally Spaced Means (ESM)

If the number of component k is quite large, the mean can be constraint to:

$$(\mu_2 - \mu_1) = (\mu_3 - \mu_2) = \cdots = (\mu_k - \mu_{k-1}). \tag{3.2.3}$$

Only two means, μ_1 and μ_2 are estimated directly. The subsequents ones are computed from the relation:

$$\mu_i = \mu_1 + (i+1)(\mu_2 - \mu_1), \qquad 3 \leq i \leq k \tag{3.2.4}$$

This assumption corresponds to linear growth in size frequency analysis and only holds if there are at least three components.

3.2.2.4 Growth Curve Means (GCM)

This constraint forces the means to lie along a von Bertalanffy growth curve of the form

$$\mu_i = L_\infty \Big\{ 1 - \exp\big[-k(t_i - t_0) \big] \Big\} \tag{3.2.5}$$

where

$$L_\infty = \mu_1 + \frac{(\mu_2 - \mu_1)^2}{(\mu_2 - \mu_1) - (\mu_3 - \mu_2)} \tag{3.2.6}$$

$$k = -log\left(\frac{\mu_3 - \mu_2}{\mu_{2-}\mu_1} \right) \tag{3.2.7}$$

$$(t_i - t_0) = -k^{-1} \log \left(1 - \frac{\mu_1}{L\infty} \right) \tag{3.2.8}$$

According to Macdonald and Green (1988), given the components (age-groups), μ_i is the mean fish size in the ith age-group, t_0 is the hypothetical age at size zero, t_i is the actual age of the ith age-group, L_∞ is the asymptotic mean size and k is the growth paramter.

Note that the growth curve constraint is only applicable if there are at least four components or age-groups and $(\mu_3 - \mu_2) < (\mu_2 - \mu_1)$ is valid.

3.2.3 Constraints on the Standard Deviations

3.2.3.1 Standard Deviaton Free (None)

The basic constraint for the standard deviations is that they must all be strictly positive. That is:

$$\sigma_i > 0, \quad (1 \leq i \leq k) \tag{3.2.9}$$

In principle, all the standard deviations can be estimated. However, it is not empirically feasible when the number of standard deviations are large.

2.3.2.3 Constant Coefficient of Variations (CCV)

With this constraint, only the first standard deviation σ_1 is estimated directly. The remaining σ_{k-1} is computed from this one and all the estimated means μ_i using the relation:

$$\sigma_i = \frac{\sigma_1}{\mu_1}\mu_i, \qquad i = 2, ..., k \tag{3.2.10}$$

The CCV constraint is also employed in the work under review.

Other constraints that could be impose on the standard deviation are **Specified Standard Deviation Fixed, all Standard Deviation Equal, and Fixed Coefficient of Variation**.

3.2.4 Identifiability

Definition 3.3

Let D be a parametric class of mixture densities and suppose further that $\forall\ g(x|\boldsymbol{\Psi}),\ g(|\boldsymbol{\Psi}^*) \in D$, then the class D is said to be identifiable for $\boldsymbol{\Psi}$ if

$$\begin{aligned} g(x|\boldsymbol{\Psi}) &= g(|\boldsymbol{\Psi}^*) \\ \Rightarrow \sum_{i=1}^{k} \pi_i f(x|\boldsymbol{\theta}_i) &= \sum_{i=j}^{k'} \pi'_j f'(x|\boldsymbol{\theta}^*_j) \end{aligned} \tag{3.2.11}$$

This further implies that $k = k^*$, and $\forall i,\ \exists j$ such that $\pi_i = \pi^*_j$ and $\boldsymbol{\theta}_i = \boldsymbol{\theta}^*_j$

The identifiability condition guarantees that there is a unique characterization of every parameter set considered, and hence no two sets of distinct parameters can give rise to the same mixture density model.

3.3 Estimating the parameters of FMM

This section discusses the various methods of estimating the parameters from FMM. Basically, there are three parametric approaches to estimating mixture parameters. These are: the graphical methods, nongraphical (statistical) methods (namely the method of moments, maximum likelihood, and various minimun distance methods), and the Bayesian method.

This paper focuses on the nongraphical methods particularly estimation by the ML method. Unfortunately, the *log* likelihood functions of most FMM and incomplete data problems often do not exist in closed forms. Hence analytically intractable to compute the parameter estimate. To solve this probem, a brief discription of how to use some iterative procedures: the *Expectation Maximization* (EM) algorithm and the *Newton Raphson* (NR) *algorithm* to approximate these solutions are examined.

3.3.1 The method of moments

Let n be identically independent observations from a k component population with densities functions and a k unknown parameter set $\boldsymbol{\theta}$. Suppose further that $\boldsymbol{\nu(\theta)}$ denote a vector of k independent moments and $\boldsymbol{m}$ the corresponding sample moments. Then the method of moments estimator $\hat{\boldsymbol{\theta}}$ satisfies the relation:

$$\boldsymbol{\nu}(\hat{\boldsymbol{\theta}}) = \boldsymbol{m} \tag{3.3.1}$$

This method is considered the least efficient and computationally unstable. Also, in practice, when the components are more than two, it is rarelly applicable.

3.3.2 The Minimum Distance Estimation Method

In this method various distance funcitons are used to compute the discrepancies that exist between the theoretical and the empirical distribution functions ob-

tained from a sample of n identically independent observations. The distance functions together with some numerical methods are then used to approximate the parameters that minimize the discrepancies. This method is usually used for binned data.

Definition: 3.4

Suppose $G(x|\boldsymbol{\Psi})$ *is the theoretical mixture distribution density function of interest with* $\boldsymbol{\Psi}$, *the unknown parameters. Suppose further that* $G_n(x)$ *is the empirical distribution function obtained from the observed sample. If* $\delta(G, G_n)$ *is the measure of the discrepancy between the functions* $G(.)$ *and* $G_n(.)$. *Then, the minimum distance estimator,* $\hat{\boldsymbol{\Psi}}$ *for* $\boldsymbol{\Psi}$ *is the value of* $\boldsymbol{\Psi}$ *that minimizes*

$$\delta[G_n(x), G(x|\boldsymbol{\Psi})] \tag{3.3.2}$$

Note that if the distance measure is $\delta(\boldsymbol{\Psi})$, then $\hat{\boldsymbol{\Psi}}$ is the critical or the stationary point (the estimator of the parameter of interest $\boldsymbol{\Psi}$) of $\delta(\boldsymbol{\Psi})$ such that:

$$\frac{\partial \delta(\hat{\boldsymbol{\Psi}})}{\partial \boldsymbol{\Psi}} = 0 \tag{3.3.3}$$

There are various estimators that can be derived from this method all depending on the type of distance function used to measure the discrepancy, and also the choice of the numerical method employ to approximate the parameter values. The choice of the measure is very important. This will determine the statistical properties of the estimates such as asymptotic consistency and continuous partial derivative with respect to the paramter of interest $\boldsymbol{\Psi}$. Also, this has a bearing on the choice of numerical methods needed for the minimization (Macdonald and Pitcher, 1979).

A few useful distance functions are listed below. Suppose the data is grouped over

k size intervals with $n_i, (i = 1, ..., k)$ being the observed size frequencies such that $\sum n_i = n$. Suppose further that $\pi_i(\boldsymbol{\theta}), (i = 1, ..., k)$ is the hypothetical propability distribution with respect to each parameter $\boldsymbol{\theta}$. Then the Chi-Square χ^2 statistic (distance function or measure) is given as

$$\begin{aligned} \delta_{\chi^2}(\boldsymbol{\theta}) &= \sum_{i=1}^{k} \frac{(n_i - n\pi_i(\boldsymbol{\theta}))^2}{n\pi_i(\boldsymbol{\theta})} \\ &= n\sum_{i=1}^{k} \frac{\left(\frac{n_i}{n} - \pi_i(\boldsymbol{\theta})\right)^2}{\pi_i(\boldsymbol{\theta})} \\ &= n\sum_{i=1}^{k} \frac{(\hat{p}_i - p_i(\boldsymbol{\theta}))^2}{p_i(\boldsymbol{\theta})}, \end{aligned} \tag{3.3.4}$$

where; $\hat{p}_i = \frac{n_i}{n}$ is the observed relative frequency of the i^{th} interval, $p_i(\boldsymbol{\theta})$ is the hypothetical probability function with respect to the parameter $\boldsymbol{\theta}$.

Similarly, the Modified Chi-Square χ^2_{mod} distance measure is given as

$$\delta_{\chi^2_{mod}}(\boldsymbol{\theta}) = \sum_{i=1}^{k} \frac{(\hat{p}_i - p_i(\boldsymbol{\theta}))^2}{\hat{p}_i} \tag{3.3.5}$$

where $0 < \hat{p}_i \leq 1$

Other distance measure functions used in the estimation of grouped data are:

Hillinger Distance given by:

$$\delta_{H.D}(\boldsymbol{\theta}) = \sum_{i=1}^{k} \left| \sqrt{\hat{p}_i} - \sqrt{p_i(\boldsymbol{\theta})} \right|^2 \tag{3.3.6}$$

The Kullback-Leibler Seperator:

$$\delta_{K.L}(\boldsymbol{\theta}) = \sum_{i=1}^{k} p_i(\boldsymbol{\theta}) \log \frac{p_i(\boldsymbol{\theta})}{\hat{p}_i} \tag{3.3.7}$$

Haldane's discrepancy:

$$D_k = \frac{(n+k)!}{n!} \sum_{i=1}^{k} \frac{n_i! p_i^{k+1}(\boldsymbol{\theta})}{(n_i + k)!}, \tag{3.3.8}$$

where $k \neq 1$. For $k = -1$, then

$$D_{-1} = -\frac{1}{n}\sum_{i=1}^{k} n_i \log p_i(\boldsymbol{\theta})$$

Note that the Kullback-Leibler distance measure $\delta_{K.L}(\boldsymbol{\theta})$ when minimized, gives exactly the same estimate $\hat{\boldsymbol{\theta}}$ as maximizing the log likelihood function

$$\log L(\boldsymbol{\theta}|\hat{\boldsymbol{p}}) = n\sum_{i=1}^{k} \hat{p} \log p_i(\boldsymbol{\theta})$$

All these distance functions when minimized give equivalent best asymptotic normal property like the ML. That is, in large sample, they are asymptotically consistent, normal, and have continuous partial derivatives like the ML. In particular, Neyman (1949) and Titterington et al. 1985 showed that the system of equations of the first partial derivatives of the Chi-square and the Modified Chi-square distance funcitons with respect to the k parameter $\boldsymbol{\theta}_i, (i = 1, .., k)$ satisfy the best asymptotic normal property of the ML in large sample. Thus, preferred in situations where the data come grouped. That is

$$\begin{aligned} \frac{\partial \delta_{\chi^2}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i} &= -n\sum_{i=1}^{k} \left(\frac{\hat{p}_i}{p_i(\boldsymbol{\theta})}\right)^2 p_i^{'}(\boldsymbol{\theta}) \\ &= 0 \end{aligned}$$

and

$$\frac{\partial \ _{\text{mod}}\chi^2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i} = \sum_{i=1}^{k} \frac{p_i(\boldsymbol{\theta})}{\hat{p}_i} p^{'}(\boldsymbol{\theta}) = 0$$

when solved with the the appropriate numerical methods deliver the same asymptotic estimate property in large sample like the maximum likelihood estimates.

In matrix notations, let $\boldsymbol{p} = \boldsymbol{p}(\boldsymbol{\theta})$ and $\hat{\boldsymbol{p}}$ be some $m - row$ vectors and $\boldsymbol{p}(\boldsymbol{\theta})$ a function of the $s-$row vector parameter. Also suppose that $\boldsymbol{V} = \boldsymbol{p}/n$ is some $m \times m$ square diagonal, symmetric and positive definite matrix with zeros off the

leading diagonals. Then $\delta_{\chi^2}(\boldsymbol{\theta})$ and $\delta_{\chi^2_{mod}}(\boldsymbol{\theta})$ can be written as:

$$\delta_{\chi^2}(\boldsymbol{\theta}) = (\hat{\boldsymbol{p}} - \boldsymbol{p})' \boldsymbol{V}^{-1} (\hat{\boldsymbol{p}} - \boldsymbol{p}) \tag{3.3.9}$$

$$\delta_{\chi^2_{mod}}(\boldsymbol{\theta}) = (\hat{\boldsymbol{p}} - \boldsymbol{p})' \hat{\boldsymbol{V}}^{-1} (\hat{\boldsymbol{p}} - \boldsymbol{p}) \tag{3.3.10}$$

respectively.

The minimization of the quadratic distance measure (χ^2 and mod .χ^2) have the following advantages over the maximization of the log likelihood funtion:

1. the minimized distance method (Chi-square) is simpler to implement.

2. the estimates are directly used in the computation of the Chi-square GOF test.

3. it is easier to interprete as the discrepancy between the theoretical and the empirical distribution.

In most empirical situations, the stationary equations from the first partial derivatives of the distance measure quadratic functions do not exist in closed form, therefore the parameters have to be approximated iteratively. In such case, the NR quadratic method is chosen.

3.3.3 The Maximum Likelihood (ML) Estimation Method

The ML estimation method is considered by far the most consistent and efficient mixture estimation method. It has desirable statistical properties. The estimates are asymptotically consistent. That is, they converge to the true parameter values if the sample size is large enough. The estimates are also consisderd to be asymptotically efficient and unbiased meaning that they are most precise and give the right value on average for large sample compared to other methods. This section reviews the general likelihood and log likelihood functions, the score/gradient statistic, the hessian, the observe information, and its equivalent expected infor-

mation matrix of the *mle*.

Definition 3.5

Let X be an iid random variable that assume values $\boldsymbol{x} \equiv (x_1, \ldots, x_n)^T$ *on a vector random sample of size n .Also, suppose* $\boldsymbol{\Psi} \equiv (\Psi_1, \ldots, \Psi_n)^T$ *is the vector of distinct parameters to be estimated on the parameter space* $\boldsymbol{\Omega}$. *If the pdf is* $f_i(x_i|\boldsymbol{\Psi})$, *then the likelihood function for the entire sample is given as*:

$$L_n(\boldsymbol{\Psi}, \boldsymbol{x}) = \prod_{\boldsymbol{i=1}}^{\boldsymbol{n}} f(x_i|\boldsymbol{\Psi}) \tag{3.3.11}$$

Instead of maximizing the likelihood, the log *likehood function*

$$logL_n(\boldsymbol{\Psi}, \boldsymbol{x}) = \ell_n(\boldsymbol{\Psi}, x) = \sum_{i=1}^{n} log\{f(x_i|\boldsymbol{\Psi})\} \tag{3.3.12}$$

is rather maximized.

Definition 3.6

The maximum likelihood estimator $\hat{\boldsymbol{\Psi}}$ *of* $\boldsymbol{\Psi}$ *is the value of* $\hat{\boldsymbol{\Psi}}$ *that maximizes* $\ell_n(\boldsymbol{\Psi}, \boldsymbol{x})$ *such that*:

$$\ell_n(\hat{\boldsymbol{\Psi}}, \boldsymbol{x}) \geq \ell_n(\boldsymbol{\Psi}, \boldsymbol{x}), \quad \forall \boldsymbol{\Psi} \in \boldsymbol{\Omega} \tag{3.3.13}$$

Note: If the *mle*, $\hat{\boldsymbol{\Psi}}$ is unique which is always not the case and $L_n(\boldsymbol{\Psi}, \boldsymbol{x})$ or $\ell_n(\boldsymbol{\Psi}, \boldsymbol{x})$ is bounded and differentiable, then the values of $\hat{\boldsymbol{\Psi}}$ can be found by using the score/gradient equation..

Definition 3.7

The score or the gradient equation is the simultaneous equations of the first partial derivatives of $\ell_n(\boldsymbol{\Psi}, x)$ *with respect to each parameter in* $\boldsymbol{\Psi}$ *such that*

$$
\begin{aligned}
S_n(\boldsymbol{x}, \boldsymbol{\Psi}) &= \frac{\partial \ell_n(\boldsymbol{\Psi}, \boldsymbol{x})}{\partial \boldsymbol{\Psi}} \\
&= \sum_{i=1}^{n} \frac{1}{f(x_i|\boldsymbol{\Psi})} \left\{ \frac{\partial}{\partial \boldsymbol{\Psi}} f(x_i|\boldsymbol{\Psi}) \right\} \\
&= 0
\end{aligned}
\tag{3.3.14}
$$

To prove that indeed, the *mle*, $\hat{\boldsymbol{\Psi}}$ is the maximum estimate but no other critical point, the **Observed Information Matrix or** negative of the **Hessian matrix** is computed.

Definition 3.8

The observed information matrix is the negative of the second order partial derivatives (the Hessian matrix) of the loglikelihood function, with respect to elements of $\boldsymbol{\Psi}$. That is:

$$
\begin{aligned}
I_n(\boldsymbol{\Psi}, \boldsymbol{x}) &= -\frac{\partial \ell_n^2(\boldsymbol{\Psi}, \boldsymbol{x})}{\partial \boldsymbol{\Psi} \partial \boldsymbol{\Psi}^T} \\
&= -H(\boldsymbol{\Psi}, \boldsymbol{x})
\end{aligned}
\tag{3.3.15}
$$

where $H(\boldsymbol{\Psi}, \boldsymbol{x}) = \dfrac{\partial \ell_n^2(\boldsymbol{\Psi}, \boldsymbol{x})}{\partial \boldsymbol{\Psi} \partial \boldsymbol{\Psi}^T}$ is the Hessian matrix.

Definition 3.9

The Fisher or the expected information matrix under regularity conditions is the expectation of the observed information matrix. Alternatively, it is the negative expectation of the Hessian matrix given by:

.

$$\begin{aligned} \boldsymbol{I}_n(\boldsymbol{\Psi}) &= E[S_n(\boldsymbol{\Psi})S_n^T(\boldsymbol{\Psi})] \\ &= E[I_n(\boldsymbol{\Psi}, \boldsymbol{x})] \\ &= -E[H(\boldsymbol{\Psi}, \boldsymbol{x})] \end{aligned} \tag{3.3.16}$$

Note that the asymptotic variance covariance matrix is the inverse of the expected information matrix. That is:

$$\begin{aligned} Var(\hat{\boldsymbol{\Psi}}) &= \boldsymbol{I}_n^{-1}(\hat{\boldsymbol{\Psi}}, x) \\ SE(\hat{\boldsymbol{\Psi}}) &= \{\boldsymbol{I}_n^{-1}(\hat{\boldsymbol{\Psi}})\}_{ii}^{\frac{1}{2}} \end{aligned} \tag{3.3.17}$$

where the standard error $SE(\hat{\boldsymbol{\Psi}})$ is the the square root of the diagonal elements of variance covariance matrix $Var(\hat{\boldsymbol{\Psi}})$.

In practice, most of the log likelihood functions are complicated and do not exist in closed-forms and hence the parameter estimation are analytically intractable. In such cases, numerical procedures such as the EM and NR algorithms are employed in the estimation process. In the case of estimating the parameters of the mixture models, the likelihood functions are complex depending on whether the problem is a continuous or of incomplete data (missing data, grouped, censored, or truncated) problem . The next subsections describe the data type used in this work. Finally, the likelihood functions of the finite mixture distributions, grouped, and combined likelihood functions of two are examined.

Definition 3.10 Complete Data Problem

Let $x_1, ..., x_n$ be n identically independent realized observations of the random vector $\boldsymbol{X}$ from a mixture density. Suppose further that $\boldsymbol{z} = (z_1, .., z_n)$ is a k dimensional components indicator vector values of zeros and once of the

vector random variable $\boldsymbol{Z}$ *such that the missing values are define as*

$$z_{ij} = \begin{cases} 1, & if\ x_i\ belongs\ to\ the\ j^{th}\ component \\ 0, & if\ x_i\ does\ not\ belong\ to\ the\ j^{th}\ component \end{cases}$$

for $i = 1, .., n$ *and* $j = 1, ..., k$, *then the complete data observation vector* $\boldsymbol{y}_c$ *from the random vector variable* $\boldsymbol{Y}$ *is given as*

$$\boldsymbol{y}_c = (\boldsymbol{x}, \boldsymbol{z}) \tag{3.3.18}$$

Here, both the individual observations and their respective subpopulations are directly observed.

For the complete data vector $\boldsymbol{y}_c$, the likelihood and log likelihood functions for the parameter $\boldsymbol{\Psi}$ are given by:

$$\begin{aligned} L_n(\boldsymbol{\Psi}, \boldsymbol{y}_c) &= \prod_{j=1}^{n} g(x_j|\boldsymbol{\Psi}) \\ &= \prod_{j=1}^{n} \left\{ \prod_{i=1}^{k} \pi_i^{z_{ij}} f(x_j|\boldsymbol{\theta}_i)^{z_{ij}} \right\} \end{aligned} \tag{3.3.19}$$

and

$$\begin{aligned} \ell_n(\boldsymbol{\Psi}, \boldsymbol{y}_c) &= \sum_{j=1}^{n}\sum_{i=1}^{k} z_{ij} log\pi_i + \sum_{j=1}^{n}\sum_{i=1}^{k} z_{ij} log f_i(x_j|\boldsymbol{\theta}_i) \\ &= \sum_{j=1}^{k}\sum_{i=1}^{n} z_{ij} \{ \log \pi_i + \log f_i(x_j, |\boldsymbol{\theta}_i) \} \end{aligned} \tag{3.3.20}$$

respectively.

Note that *mle* of the log likelihood function of the complete data problem can be obtained directly under some assumptions.

Most of the statistical problems fall under the category of incomplete data problems and the finite mixture densities are no exceptions. Unfortunately, the log likelihoods of these densities are quite complicated and hence the ML not appli-

cable directly. The challenge then is to transform the incomplete data problem to a complete one by introducing the appropriate missing variable.

Definition 3.11 Grouped Data

Let X be a random variable with density function $f(x;\boldsymbol{\Psi})$. Suppose further that the sample space of X is partioned into m exclusive interval with boundaries $a_0,\ldots,a_n$. Individual observations $x_1,\ldots,x_n$ are made on X but only the number of observation (frequency) n_i falling into interval $[a_{i-1},a_i)_{i=1,\ldots,m}$ are reported such that $n=\sum_{i=1}^{m} n_i$. Then the data form:

$$y_i=(a_i,n_i),\quad i=1,\ldots,m \tag{3.3.21}$$

is call a grouped data.

Note that the observed data vector $\boldsymbol{y}=(n_1,\ldots,n_m)$ has multinomial distribution consisting of n draws on m categories with probabilities $P_1(\Psi),\ldots P_m(\Psi)$, where

$$P_i(\boldsymbol{\Psi})=\int_{a_{i-1}}^{a_i} f(x,\boldsymbol{\Psi})d\boldsymbol{\Psi},\qquad i=1,\ldots m \tag{3.3.22}$$

is the probability that an individual x belongs to the ith interval and $f(x,\boldsymbol{\Psi})$ is the density function of X.

Note that the likelihood function of the observed data $\boldsymbol{y}$ with multinomial probability $P_i(\boldsymbol{\Psi})$ is given as:

$$L(\boldsymbol{\Psi})=\frac{n!}{n_1\ldots n_m}[P_1(\boldsymbol{\Psi})]^{n_1}\ldots[P_m(\boldsymbol{\Psi})]^{n_m} \tag{3.3.23}$$

In computing the log likelihood, only the part of the likelihood function that depend on the parameter of interest $\boldsymbol{\Psi}$ is used. That is:

$$logL(\boldsymbol{\Psi})=\ell(\boldsymbol{\Psi})=\sum_{i=1}^{m} n_i\log P_i(\boldsymbol{\Psi}) \tag{3.3.24}$$

The next subsection deals with the general theory behind the $EM-algorithm$. This algorithm is used to approximate the finite mixture grouped parameters. Finally, the $Newton\ Raphson$ (NR) numerical method which is used in the $M-step$ of the $EM-algorithm$ if the function from the E-step is not in a close form as discussed.

3.3.4 The Expectation Maximization (EM) Algorithm

The $EM\ Algorithm$ is one of the commonly iterative method used in approximating ML mixture parameters from complex data problems. Common and less complicated approach is its application to continuous data. In this work, the EM algorithm is used to approximate the GFMM parameters given that the data come grouped. The log likelihood of these incomplete data do not exist in close form and hence the parameter estimates are analytically intractable.

The EM algorithm consists of two steps: the $Expectation\,step$, (E-step) and the $Maximization-step$ (M-step). In the E-step, the incomplete observed data is transformed to a complete one whose log likelihood function often exist in close form, and is easy to solve analytically. Instead of the incomplete data, the conditional expectation of the complete data given the observed data is rather use. Together the appropriate initial values, the E-step generate some function $Q(.)$ which is iteratively maximized in the M-step until convergence or stopping critirion is reached.

Summary of the EM-algorithm

1. Transform the given incomplete data to complete data
2. generate a certain $Q(.)$ function consisting of the conditional expectation of the complete data log likelihood given that the observe exist and obtain appropriate initial values.
3. maximize the $Q(.)$ function in the m-step to obtain new estimates

4. update the initial values with the new estimates.

5. repeat step 2 and 3 until convergence and/or stopping critirion is reached.

Mathematically, the EM Algorithms can be formulated as follows:

Let $\boldsymbol{Y}$ be random vector variable assuming observed incomplete data $\boldsymbol{y} = (y_i, \ldots, y_n)^T$ with *pdf* $g(\boldsymbol{y}|\boldsymbol{\Psi})$, where $\boldsymbol{\Psi}$ is the vector of parameters to be estimated. Suppose further that $\boldsymbol{X}$ is the random vector variable assuming the complete data values $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ with *pdf* $f_c(\boldsymbol{x}|\boldsymbol{\Psi})$ and log likelihood $log\, f_c(\boldsymbol{x}|\boldsymbol{\Psi})$. Let $\boldsymbol{\Psi}^{(0)}$ be the appropriate initial value of $\boldsymbol{\Psi}$. Then on the first iteration, the E-step requires the calculation of the conditional expectation of the complete data log likehood given the observed data x. That is

$$E_{\boldsymbol{\Psi}^{(0)}}[\log L_c(\boldsymbol{\Psi})\,|X] = Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(0)}) \tag{3.3.25}$$

where $\boldsymbol{\Psi}^{(0)}$ is the initial value.

In the M-step, $Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(0)})$ is maximized with respect to $\boldsymbol{\Psi}$ over the parameter space $\boldsymbol{\Omega}$. That is, choose each $\boldsymbol{\Psi}^{(1)}$, $\forall \boldsymbol{\Psi} \in \boldsymbol{\Omega}$ such that

$$\boldsymbol{\Psi}^{(1)} = \arg \max_{\boldsymbol{\Psi}} [Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(0)}] \tag{3.3.26}$$

The E-and the M-step are then carried out repeatedly but at this time, $\boldsymbol{\Psi}^{(0)}$ is replaced by $\boldsymbol{\Psi}^{(1)}$.

On the $(k+1)$th iteration, the E- and M-step is given as:

E-step: $Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(k)}) = E_{\boldsymbol{\Psi}^{(k)}}[\log L_c(\boldsymbol{\Psi})|X = x]$

M-step: $\boldsymbol{\Psi}^{(k+1)} = \arg \max_{\boldsymbol{\Psi}} [Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(k)})]$

This process is alternated repeatedly until a certain stopping critirion say TOL (ε) or convergence is attained.

$$||\log L(\boldsymbol{\Psi}^{(k+1)}, x) - \log L(\boldsymbol{\Psi}^{(k)}, x)|| \leq \varepsilon, \tag{3.3.27}$$

where ε is some small chosen value.

That is, the iteration terminate at the $(k+1)th$ step if the absolute difference in the log likelihood of $\log L(\boldsymbol{\Psi}^{(k+1)}, x) - \log L(\boldsymbol{\Psi}^{(k)}, x)$ is smaller than the chosen ε. Note that alternatively, the stopping critirion could also be calculated using the change in the parameters after each iteration instead of the log likelihood. Also, convergence is sensitive to different initial values and stopping critirion.

3.3.5 Monotonicity Property of the EM Algorithm.

It is one of the major properties that gurantees convergence of the EM-algorithm. This theorem ensure that as the EM-algorithm iterates, the $(k+1)$th guess of $\boldsymbol{\Psi}^{(k+1)}$will never be less likely (perform worse) than the kth guess $\boldsymbol{\Psi}^{(k)}$ in terms of their log likelihoods.

Theorem 3.1

Let $\log g(\boldsymbol{y}|\boldsymbol{\Psi})$ be the log likelihood function. For $\boldsymbol{\Psi} \in \Omega$,
if $Q(\boldsymbol{\Psi}^{(k+1)}|\boldsymbol{\Psi}^{(k)}) > Q(\boldsymbol{\Psi}^{(k)}|\boldsymbol{\Psi}^{(k)})$, then

$$\log L(\boldsymbol{\Psi}^{(k+1)} \geq \log L(\boldsymbol{\Psi}^{(k)})$$

3.4 Estimating GFMM parameters from Grouped Data via EM Algorithm

Here, two distinct incomplete data problems with different log likelihoods (finite mxiture and grouped data) is combined into one complete data log likelihood function. Fitting GFMM to the combine data log likelihood function through EM algorithm takes two different approaches depending on whether individual observations are reported or the data come grouped. In practice, the data are grouped in the form of a histogram. Everitt et al.(2000) and McLachlan, et al

(2000) have showed situations in which mixtures were fitted to continuous data.

Definition 3.12

Let X be a random variable with mixture density function $g(x|\boldsymbol{\Psi}) = \sum_{i=1}^{k} \pi_i f(x|\boldsymbol{\theta}_i)$, where $\boldsymbol{\Psi}$ is the paramter vector in the space $\boldsymbol{\Omega}$. If $y = (n_1, \ldots, n_m)$ is the number of observations falling into the intervals $[a_0, a_1), [a_1, a_2), \ldots, [a_{m-1}, a_m)$ with boundaries $a_0, a_1, \ldots, a_m$ such that instead of the individual observations, n_i (the frequency) are rather reported. Then, the log likelihood of the GFMM given a grouped data is:

$$\log L(\boldsymbol{\Psi}) = \ell(\boldsymbol{\Psi}) = \sum_{i=1}^{m} n_i \log P_i(\boldsymbol{\Psi}) \tag{3.4.1}$$

where

$$\begin{aligned} P_i(\boldsymbol{\Psi}) &= \int_{a_{i-1}}^{a_i} g(x|\boldsymbol{\Psi}) dx \\ &= \int_{a_{i-1}}^{a_i} \{\sum_{j=1}^{k} \pi_j f(x|\boldsymbol{\theta}_j)\} dx \\ &= \sum_{j=1}^{k} \pi_j \int_{a_{i-1}}^{a_i} f(x|\boldsymbol{\theta}_j) dx \\ &= \sum_{j=1}^{k} \pi_j P_i(\boldsymbol{\theta}_j) \end{aligned} \tag{3.4.2}$$

Therefore the log likelihood function of the grouped data can be written as:

$$\begin{aligned} \log L(\boldsymbol{\Psi}) &= \sum_{i=1}^{m} n_i \log \sum_{j=1}^{k} \pi_j P_i(\boldsymbol{\theta}_j) \\ &= \sum_{i=1}^{m} \sum_{j=1}^{k} n_i \log \pi_j + \sum_{i=1}^{m} \sum_{j=1}^{k} n_i \log P_i(\boldsymbol{\theta}_j) \end{aligned} \tag{3.4.3}$$

Note that the log likelihood in (3.4.3) is not in a close form, hence can not be solved explicitely. Macdonald and Pitcher (1979) proposed two iterative procedures: the EM and NR algorithm to approximate these parameters at pre-define constraints. A missing data variable $(n^*_{ij})_{i=1,\ldots,m \; j=1,\ldots,k}$, is introduced, which is the number of observations from the jth group falling into the ith interval. That

is:

$$\boldsymbol{n}_i^* = (n_{1i}^*, \ldots, n_{ki}^*)^T, \quad i = 1, \ldots, m$$

The new complete data log likelihood $w_i = (y_i, n_i^*)$ can then be written as:

$$\begin{aligned} \log L_c(\boldsymbol{\Psi}, w) &= \sum_{i=1}^{m}\sum_{j=1}^{k} n_{ij}^* \log \pi_j + \sum_{i=1}^{m}\sum_{j=1}^{k} n_{ij}^* \log P_{ij}(\boldsymbol{\theta}_j) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{k} n_{ij}^* \log [\pi_i P_{ij}(\boldsymbol{\theta}_j)] \end{aligned} \tag{3.4.4}$$

The **E-step** for the $\log L_c(\boldsymbol{\Psi}, w)$ above is computed as:

$$\begin{aligned} Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(s)}) &= E_{\boldsymbol{\Psi}^{(s)}}\left[\log L_c(\boldsymbol{\Psi}, W)|Y = y\right] \\ &= E_{\boldsymbol{\Psi}^{(s)}}[\sum_{i=1}^{m}\sum_{j=1}^{k} N_{ij}^* \log\left[\pi P_{ij}(\boldsymbol{\theta}_j)|y\right] \\ &= \sum_{i=1}^{m}\sum_{j=1}^{k} \log \pi_i P_{ij}(\boldsymbol{\theta}_j) E_{\boldsymbol{\Psi}^{(k)}}\left[N_{ij}^*|y\right] \end{aligned} \tag{3.4.5}$$

where $E_{\boldsymbol{\Psi}^{(k)}}\left[N_{ij}^*|y\right]$ is the conditional expected value of n_{ij}^* given that the observed value y is calculated as:

$$\begin{aligned} E_{\boldsymbol{\Psi}^{(s)}}[N_{ij}^*|y] &= P(N_{ij}^* = 1|Y_i = y_i) \\ &= \frac{P(Y_i = y_i|N_{ij}^* = 1)P(N_{ij}^* = 1)}{\sum_{j=1}^{k} P(Y_i = y_i|N_{ij}^* = 1)P(N_{ij}^* = 1)} \\ &= n_i \frac{\pi_j^{(s)} P_{ij}(\boldsymbol{\theta}_j^{(s)})}{\sum_{j=1}^{k} \pi_j^{(s)} P_{ij}(\boldsymbol{\theta}_j^{(s)})} \\ &= n_i \frac{\pi_j^{(s)} f(x, \boldsymbol{\theta}_j)}{g(x_i, \boldsymbol{\Psi}^{(s)})} \\ &= e_{ij}^{*(s)} \end{aligned} \tag{3.4.6}$$

Substituting (3.4.6) into (3.4.5), the $Q(.)$ function can be computed as:

$$\begin{aligned} Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(s)}) &= \sum_{i=1}^{m}\sum_{j=1}^{k} \log \pi_i P_{ij}(\boldsymbol{\theta}_j).e_{ij}^{*(s)} \\ &= \sum_{i=1}^{m}\sum_{j=1}^{k} e_{ij}^{*(s)} \log \pi_j + \sum_{i=1}^{m}\sum_{j=1}^{k} e_{ij}^{*(s)} \log f(x_i|\boldsymbol{\theta}_j) \end{aligned} \tag{3.4.7}$$

The **M-step:**

In this step, the $Q(.)$ function generated from the **E-step** is maximized with respect to π_j and $\boldsymbol{\theta}_j$ independently since they are not related. To write the gradient equation for π_j, the Lagrange multiplier λ is introduced and the constraint $\sum_j^k \pi_j = 1$ is imposed.

$$\frac{\partial}{\partial \pi_j}\left[\sum_{i=1}^{m}\sum_{j=1}^{k}\log \pi_j e_{ij}^{*(s)} + \lambda\Big(\sum_{j=1}^{k}\pi_j - 1\Big)\right] = 0 \tag{3.4.8}$$

or

$$\sum_{i=1}^{m}\frac{1}{\pi_j}e_{ij}^{*(s)} + \lambda = 0 \tag{3.4.9}$$

summing both sides over j, then $\lambda = -m$. Therefore $\pi_j = \frac{1}{m}\sum_{i=1}^{m} e^{*(s)}$ and the iterative estimator of π_j is

$$\pi_j^{(s+1)} = \frac{1}{m}\sum_{i=1}^{m} e^{*(s)} \tag{3.4.10}$$

The maximization of $Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(k)})$ with respect to $\boldsymbol{\theta}_j = (\mu_j, \sigma_j)$ independently does not exist in closed form. Therefore the estimates of $\boldsymbol{\theta}_j$ can not be obtained analytically. The NR algorithm is chosen to maximize $Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(k)})$ with respect to $\boldsymbol{\theta}_j = (\mu_j, \sigma_j)$ as implemented in the R package "*mixdist*". There are four other methods for maximizing the Q function besides the combined EM and NR algorithm implemented in mixdist by Macdonald and Du (2011). Two of these methods entail the idea of mapping or transfoming the grouped data into continuous/individual data for each interval making use of the midpoint. The functions generated from the E-step are in closed form and therefore can easily be maximize in the M-step. Other methods entail approximating $f(x_i|\boldsymbol{\theta}_j)$ by $h.f(\bar{a}, \boldsymbol{\theta}_j)$, where h is the equal width of each interval and $\bar{a}$ is the midpoint of each interval.

3.5 The Convergence Rate Matrix of the EM algorithm

The EM algorithm considered in this work implicitely defines a mapping $\boldsymbol{\Psi} \longrightarrow \boldsymbol{M}(\boldsymbol{\Psi})$ from the parameter space $\boldsymbol{\Omega}$ such that

$$\boldsymbol{\Psi}^{(s+1)} = \boldsymbol{M}(\boldsymbol{\Psi}^{(s)}), \qquad s = 0, 1, \ldots$$

If $\boldsymbol{\Psi}^{(s)}$ converges to some fixed point $\boldsymbol{\Psi}^*$, and $\boldsymbol{M}(\boldsymbol{\Psi})$ is continuous, then $\boldsymbol{\Psi}^*$ must satisfy the relation:

$$\boldsymbol{\Psi}^* = \boldsymbol{M}(\boldsymbol{\Psi}^*)$$

By Taylor series expansion of $\boldsymbol{\Psi}^{(s+1)} = \boldsymbol{M}(\boldsymbol{\Psi}^{(s)})$ in the neighbourhood of $\boldsymbol{\Psi}^*$, the approximate equation is:

$$\boldsymbol{\Psi}^{(s+1)} - \boldsymbol{\Psi}^* \approx (\boldsymbol{\Psi}^{(s)} - \boldsymbol{\Psi}^*)\boldsymbol{J}(\boldsymbol{\Psi}^*), \tag{3.5.1}$$

where

$$\boldsymbol{DM}(\boldsymbol{\Psi}) = \left(\frac{\partial \boldsymbol{M}_j(\boldsymbol{\Psi})}{\partial \boldsymbol{\Psi}_i} \right) \Bigg|_{\boldsymbol{\Psi}=\boldsymbol{\Psi}^*}$$

is some $d \times d$ Jacobian Matrix for $\mathbf{M}(\boldsymbol{\Psi}) = (M_1(\boldsymbol{\Psi}), \ldots, M_d(\boldsymbol{\Psi}))$ evaluated at $\boldsymbol{\Psi} = \boldsymbol{\Psi}^*$. Thus, in the neighbourhood of $\boldsymbol{\Psi}^*$, the EM algorithm have a linear convergence rate with a rate matrix $\boldsymbol{DM}$.

3.6 Estimating the Variance-Covariance Matrix of $(\boldsymbol{\Psi} - \hat{\boldsymbol{\Psi}})$

One of the main criticisms of the EM-algorithm in practice is that, it does not authomatically estimate the asymptotic variance-covariance matrix (from which the standard errors will be computed) for all the estimated parameters as compared

to other numerical methods like the NR-algorithm. This section will examine the method and procedures for approximating the variance-covariance matrix using the EM algorithm.

Suppose that the observed information matrix $\boldsymbol{I}_o(\boldsymbol{\Psi}|.)$ of $\boldsymbol{\Psi}$ given the variables of the observed incomplete data X and complete data X_c are given by:

$$\boldsymbol{I}_o(\boldsymbol{\Psi}|X) = -\frac{\partial^2 \log g(X|\boldsymbol{\Psi})}{\partial \boldsymbol{\Psi} \partial \boldsymbol{\Psi}^T}$$

and

$$\boldsymbol{I}_o(\boldsymbol{\Psi}|X_c) = -\frac{\partial^2 \log g(X_c|\boldsymbol{\Psi})}{\partial \boldsymbol{\Psi} \partial \boldsymbol{\Psi}^T}$$

respectively, then the variance covariance matrix $Var(\hat{\boldsymbol{\Psi}})$ given the observed variable X can be written as

$$Var(\hat{\boldsymbol{\Psi}}) = \boldsymbol{I}_o^{-1}(\hat{\boldsymbol{\Psi}}|X) \tag{3.6.1}$$

Note that the observed information matrix of the observed incomplete variable X is analytically intractable to evaluate directly, whereas that of the complete data variable X_c given by the conditional expectation on the missing variable is a very simple function that can easily be evaluated at $\hat{\boldsymbol{\Psi}}$. That is

$$\boldsymbol{I}_{oc} = E_{\boldsymbol{\Psi}}[\boldsymbol{I}_o(\boldsymbol{\Psi}, X_c)|X] \Big|_{\boldsymbol{\Psi}=\hat{\boldsymbol{\Psi}}} \tag{3.6.2}$$

Now, from the theory that the density of the complete data is the product of the densities of the observed and missing data given the observe. That is:

$$g(X_c, \boldsymbol{\Psi}) = g(X, \boldsymbol{\Psi}) g(X_{mis}|X, \boldsymbol{\Psi})$$

Writing the log likelihood of $\boldsymbol{\Psi}$ given the observed in terms of the log likelihood

of $\boldsymbol{\Psi}$ given the complete and the missing variables, generate the equation:

$$\ell(\boldsymbol{\Psi}|X) = \ell(\boldsymbol{\Psi}|X_c) - \log g(X_{mis}|X, \boldsymbol{\Psi}) \tag{3.6.3}$$

Taking the negative second partial derivative of (3.6.3), averaging over $g(X_{mis}|X, \boldsymbol{\Psi})$, and evaluating at $\boldsymbol{\Psi} = \hat{\boldsymbol{\Psi}}$, result in the equation:

$$\boldsymbol{I}_o(\hat{\boldsymbol{\Psi}}, X) = \boldsymbol{I}_{oc} - \boldsymbol{I}_{om}, \tag{3.6.4}$$

where the matrix

$$\boldsymbol{I}_{om} = -E_{\boldsymbol{\Psi}}\left(\frac{\partial^2 \log g(X_{mis}|X, \boldsymbol{\Psi})}{\partial \boldsymbol{\Psi} \partial \boldsymbol{\Psi}^T}\middle| X, \boldsymbol{\Psi}\right)\Bigg|_{\boldsymbol{\Psi}=\hat{\boldsymbol{\Psi}}} \tag{3.6.5}$$

is the missing information from Orchard et al. (1972) missing information principle. From (3.6.4), the observed information matrix is the difference between the complete information and the missing information matrix.

Rewritting (3.6.4) as

$$\boldsymbol{I}_o(\hat{\boldsymbol{\Psi}}, X) = (\boldsymbol{I} - \boldsymbol{I}_{om}\boldsymbol{I}_{oc}^{-1})\boldsymbol{I}_{oc} \tag{3.6.6}$$

and setting $\boldsymbol{DM} = \boldsymbol{I}_{om}\boldsymbol{I}_{oc}^{-1}$, the variance covariance matrix $Var(\boldsymbol{\Psi})$ can be computed as:

$$\begin{aligned} Var(\hat{\boldsymbol{\Psi}}) &= \boldsymbol{I}_{oc}^{-1}(\boldsymbol{I} - \boldsymbol{DM})^{-1} \\ &= \boldsymbol{I}_{oc}^{-1} + \boldsymbol{I}_{oc}^{-1}\boldsymbol{DM}(\boldsymbol{I} - \boldsymbol{DM})^{-1} \\ &= \boldsymbol{I}_{oc}^{-1} + \Delta Var \end{aligned} \tag{3.6.7}$$

where

$$\Delta Var = \boldsymbol{I}_{oc}^{-1}\boldsymbol{DM}(\boldsymbol{I} - \boldsymbol{DM})^{-1} \tag{3.6.8}$$

is the increase in variance due to the missing information and $\boldsymbol{I}$ is $d \times d$ identity matrix.

3.7 Newton-Raphson (NR) Method

The Newton-Raphson Method (and it variants: quasi-Newton Methods and modified Newton-Methods) is one of the most popular numerical techniques for finding the parameter estimates of the score or gradient functions in situations where the maximized log likelihood can not be solved analytically by direct ML and EM-algorithms. This algorithm is usually applied to complicated score functions/statistics by approximating it by Taylor series expansion to compute the parameter estimates numerically.

Definition: 3.13 Newton Raphson algorithm

Suppose $\boldsymbol{S}_n \in C^2[a,b]$ is a vector function of the set of all functions that have second continuous partial derivatives on the closed interval $[a,b]$. Let $\boldsymbol{\Psi}^{(k)} \in [a,b]$ be an approximation to $\boldsymbol{\Psi}^{(k+1)}$ such that $\boldsymbol{S}_n^{'}(x_i|\boldsymbol{\Psi}^{(k)}) \neq 0$ and also $|\boldsymbol{\Psi} - \boldsymbol{\Psi}^{(k)}|$ is very small. Then the Taylor series expansion of $\boldsymbol{S}_n(x_i|\boldsymbol{\Psi}^{(k+1)})$ about $\boldsymbol{\Psi}^{(k)}$ is

$$\begin{aligned} \boldsymbol{S}_n(x_i|\boldsymbol{\Psi}^{(k+1)}) &= \boldsymbol{S}_n(x_i|\boldsymbol{\Psi}^{(k)}) + (\boldsymbol{\Psi}^{(k+1)} - \boldsymbol{\Psi}^{k})\boldsymbol{S}_n^{'}(x_i|\boldsymbol{\Psi}^{(k)}) \\ &+ \frac{(\boldsymbol{\Psi}^{(k+1)} - \boldsymbol{\Psi}^{(k)})^2}{2!}\boldsymbol{S}_n^{''}(x_i|\xi(\boldsymbol{\Psi}^{(k+1)})) \end{aligned} \tag{3.7.1}$$

where $\xi(\boldsymbol{\Psi}^{(k+1)})$ is between $\boldsymbol{\Psi}^{(k)}$ and $\boldsymbol{\Psi}^{(k+1)}$. Now let $\boldsymbol{S}_n(x_i|\boldsymbol{\Psi}^{(k+1)}) = 0$ and $\boldsymbol{S}_n^{'}(x_i|\boldsymbol{\Psi}^{(k)}) = -\boldsymbol{I}_n(\boldsymbol{\Psi}^{(k)})$. Assuming further that $|\boldsymbol{\Psi}^{(k+1)} - \boldsymbol{\Psi}^{(k)}|$ is small, then the third term becomes insignificantly smaller and hence can be dropped and solving for $\boldsymbol{\Psi}^{(k+1)}$ in (3.7.1), results in the equation:

$$\boldsymbol{\Psi}^{(k+1)} = \boldsymbol{\Psi}^{(k)} + \boldsymbol{I}_n^{-1}(x_i|\boldsymbol{\Psi}^{(k)})\boldsymbol{S}_n(x_i|\boldsymbol{\Psi}^{(k)}), \tag{3.7.2}$$

where $\boldsymbol{I}_n(.)$ is the information matrix and $\boldsymbol{S}_n(.)$ the gradient vector statistic. In

matrix notation, let

$$\begin{aligned}\boldsymbol{D} = & \frac{\partial}{\partial \boldsymbol{\Psi}} \delta[G_n(x), G(x|\boldsymbol{\Psi})] \\ = & \begin{bmatrix} \frac{\partial p_1}{\partial \Psi_1}(\boldsymbol{x}) & \cdots & \frac{\partial p}{\partial \Psi_s}(\boldsymbol{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial p_m}{\partial \Psi_1}(\boldsymbol{x}) & \cdots & \frac{\partial p_m}{\partial \Psi_s}(\boldsymbol{x}) \end{bmatrix}\end{aligned}$$

denote some $m \times s$ first partial derivative matrix with respect to each parameter $\boldsymbol{\Psi}$. Then $\boldsymbol{I}_n(.)$ and $\boldsymbol{S}_n(.)$ can be written as:

$$\boldsymbol{I}_n = \boldsymbol{D}' \boldsymbol{V}^{-1} \boldsymbol{D} \tag{3.7.3}$$

and

$$\boldsymbol{S} = \boldsymbol{D}' \boldsymbol{V}^{-1} \hat{\boldsymbol{p}} \tag{3.7.4}$$

The minimum distance measure (χ^2 and $\mod .\chi^2$) estimator $\hat{\boldsymbol{\Psi}}$ is the root of the non-linear normal vector equation $\boldsymbol{S} = \boldsymbol{0}$ which can only be solved numerically. That is:

$$\boldsymbol{D}' \boldsymbol{V}^{-1} \hat{\boldsymbol{p}} = \boldsymbol{0} \tag{3.7.5}$$

Applying NR numerical algorithm to (3.7.4), the new estimate can be obtain from:

$$\begin{aligned}\boldsymbol{\Psi}^1 = & \boldsymbol{\Psi}^0 + \boldsymbol{I}^{-1} \boldsymbol{S} \\ = & \boldsymbol{\Psi}^0 + (\boldsymbol{D}' \boldsymbol{V}^{-1} \boldsymbol{D})^{-1} \boldsymbol{D}' \boldsymbol{V}^{-1} \hat{\boldsymbol{p}}\end{aligned} \tag{3.7.6}$$

where, $\boldsymbol{\Psi}^0$ is the appropriate initial value for the parameter of interest $\boldsymbol{\Psi}$.

Note: In theory, Newton-Raphson method converges quadratically if and only if reasonable assumptions are made on $L_n(\boldsymbol{\Psi})$ and sufficiently accurate starting values are chosen. But in practice, the initial values are only guess and successive approximations are generated by the Newton-Raphson method. The NR-algorithm

have a quadratic rate of convergence, hence faster than the EM-algorithm. Also, it authomatically estimate the covariance matrix unlike EM-algorithm. However, it is very difficult to implement in practice due to the initial computation of some $d \times d$ inverse information matrx $\boldsymbol{I}_n^{-1}(x_i|\boldsymbol{\Psi}^k)$ for each k iteration. In addition, the initial values are very sensitive and it takes a lot of storage space.

Chapter 4

DATA COLLECTION, ANALYSIS AND RESULTS

4.0 Introduction

This chapter illustrates the use of the ML method via the EM and the NR algorithms to approximate all the mixture parameters of the yellowfin tuna population from GFMM. The form and the summary of the dataset used is also examined. The histogram is chosen to summarize the length-frequency data over other methods such as proportional distribution structure (PPS) since modes/components are clearly visible to researchers with even less statistical experience. Various stages in the analysis involves: choosing of appropriate initial values for the EM algorithms and superimposing the distributions on the histogram. The sensitivity of the fits are examined intuitively and objectively (by computing the full covariance matrix). Finally, the Chi-square GOF test is used empirically to determine whether the GFMM fits the data well. The analysis was carried out by Macdonald and Du (2011) R mixdist package.

4.1 The Nature of Data

The dataset consisted of the 2014 monthly length-frequency data of the yellowfin tuna from the log files of the Fisheries Commission of Ghana. This year was chosen because the sample size was large enough and consistent. The fork-length of juvenile tuna was measured in length (cm) and constraint to ($38cm < FL < 80cm$) interval. This size was chosen in order to avoid excessive overlapping in the components when the length are quite large. Each sample observation was first grouped into fork length, FL (in cm) with equal class intervals or width of $h = 2cm$. The histogram in figure 4.1.1 shows the summary of the quarterly grouped data for the year 2014. From the figure, it is evident that the components are heavily overlapped and do not form distinct modes.

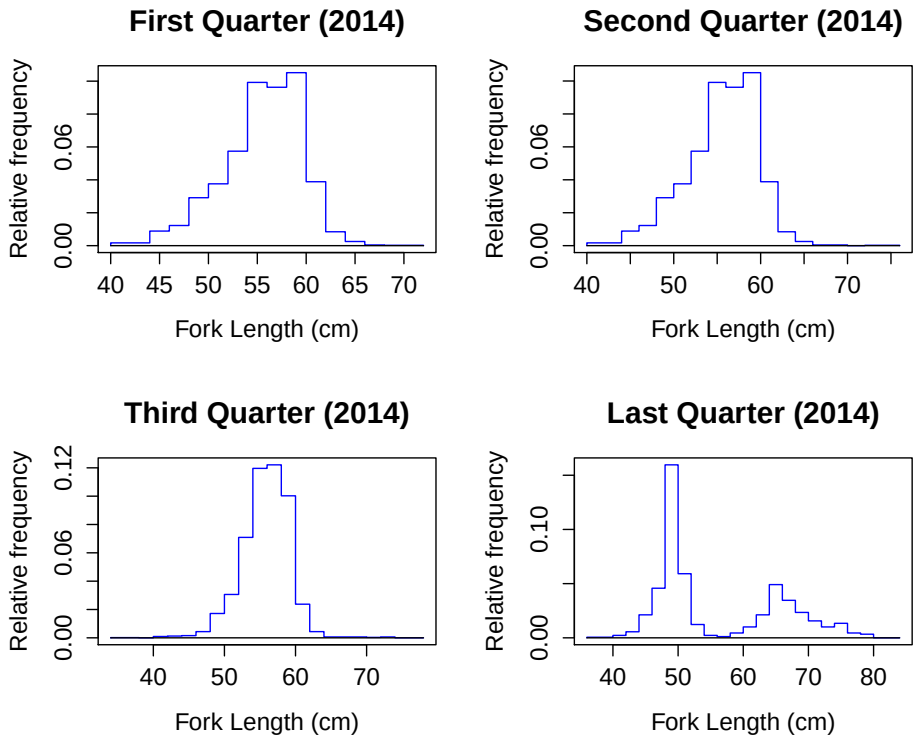

Figure 4.1.1: Histogram of quarterly length frequency of the yellowfin tuna in the year 2014

Now, in order to accurately fit the mixture distribution model to this data, initial values of all the parameters including the number of components k need to be known. Unfortunately, the correct number of components k cannot be estimated

from the histogram alone since the modes are not clearly distinct. Hence in this work, the number of components k was guessed through trial and error method aided by the histogram.

Based on the particular value of k, the initial values of the means (mu), the proportions (pi), and the standard deviations ($sigma$) were also guessed aided by visual observation from the histogram.

Note: The following constraints were imposed on the proportions and the standard deviations: for the the proportions, $0 \leq pi \leq 1$ ($i = 1, \ldots, k$); $pi_1 + \cdots + pi_k = 1$ while for the standard deviations, the CCV was imposed.

Next, in order to assess the sensitivity of these initial values, and the normality assumption, the densities of each component were superimposed on the histogram using the initial values for each components and the fits observed intuitively. Figure 4.1.2 shows that the model at least graphically fit the data well.

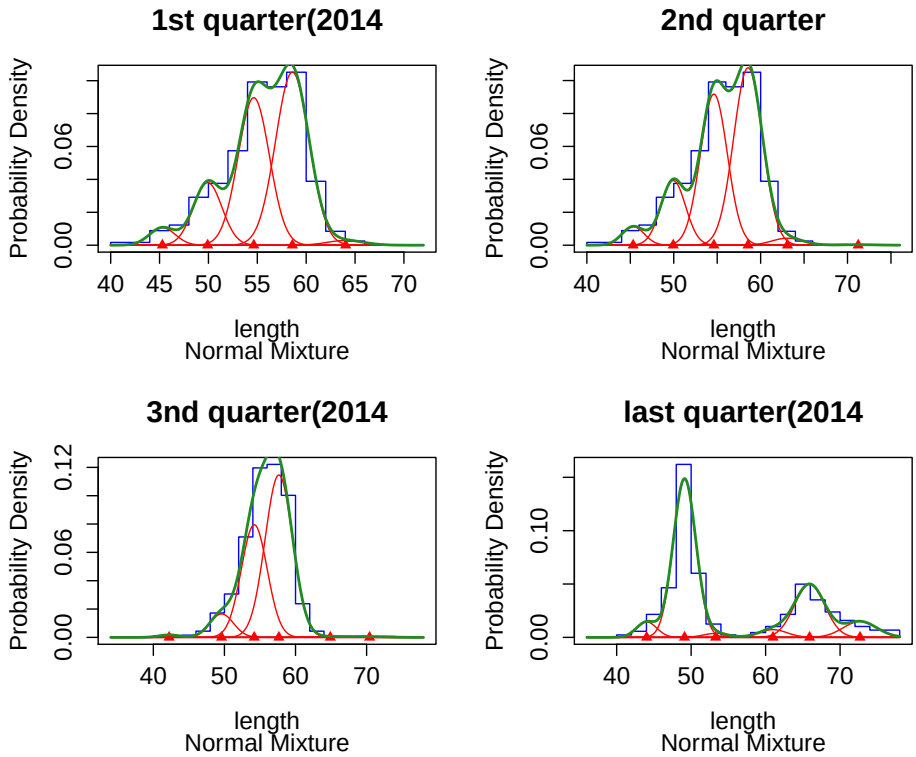

Figure 4.1.2: Quarterly length frequency histogram of 2014 yellowfin tuna fitted to GFMM

Note: the green (thick) curve represents the total fit to the GFMM whereas the

red (thin) curves represent the best fit for each component distributions from the sample with the triangle on the x-axis denoting the mean lengths positions.

The hanging rootogram was used to assess the GOF intuitively (graphically) between the the observed and the assumed GFMM. The small rectangles around the x-axis in Figure 4.1.3 show that there are no clear patterns of overfitting or/and underfitting. Hence, it can be concluded that the model fit the data well from the rootogram analyis.

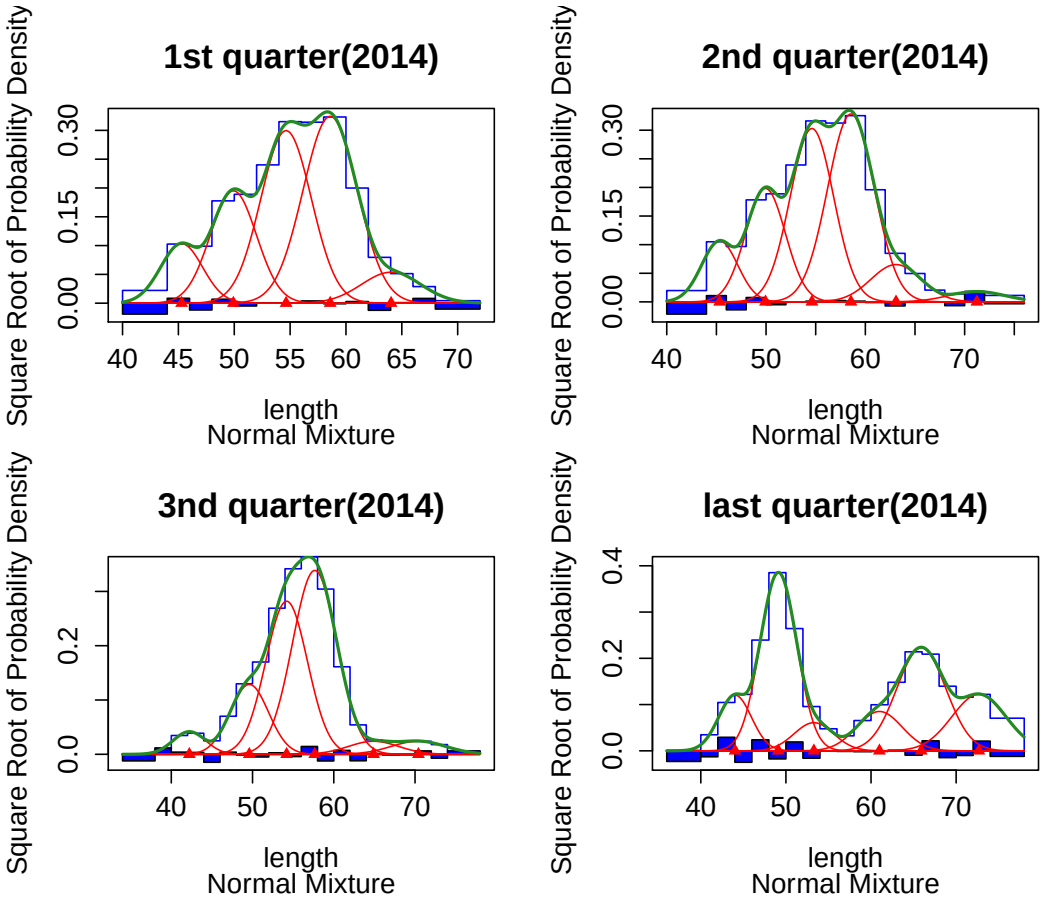

Figure 4.1.3: Assessing the fit of the model from the hanging rootogram analysis

Table 4.1 , 4.2, and 4.3 display the components with their parameter estimates, respective standard errors and Chi-square statistic respectively for the first quarter in 2014.

The proportions or weights for the first four age-groups increased gradually until the fifth age-group where it began to decrease. The continuous increament in the mean-lengths is at least twice that of the standard deviations across all the five

age-groups.

Note that only the first standard error (*sigma.se*) of the standard deviation was displayed, the rest set to NA. This is due to the fact that only one degree of freedom was used to estimate all the sigmas when the constraint CCV was imposed otherwise too many parameters will have to be estimated. The remaining standard errors could be computed from the Delta method by using the first estimated standard deviation together with all the estimated $mean-lengths$. The square root of the diagonal elements of the complete covariance matrix output all these standard errors.

In all, the total number of free parameters fitted was 10. Four free mixing proportions (*the pi add one*), one free sigma (*CCV was used*) and five free means. Hence the dimension of the inverted information matrix is (10×10). The complete covariance matrix for all the parameters estimated when the delta method was applied is captured in appendix B. Table 4. display only the standard errors computed from the square roots of the diagonal elements of the covariance matrix. The new dimension of the complete covariance matrix is now 15×15.

Table 4.4 illustrates the use of the analysis of variance ($ANOVA$) and Chi-square to objectively test the GOF of GFMM to the sample.

The Chi-square test of GOF with a $p-value$ of 0.193 shows that the GFMM fit the 2014 yellowfin tuna length frequency data in the first quarter very well.

Table 4.1: Estimated parameters with number of estimates in each group for the first quarter (2014)

age-groups (k)	proportions (pi)	mean lengths (mu)	standard deviations (sigma)	Estimated no.in group
1	0.03540	45.29	1.326	42
2	0.13932	49.92	1.461	165
3	0.35952	54.64	1.600	425
4	0.45219	58.59	1.715	536
5	0.01357	64.04	1.875	16

Table 4.2: Estimated parameters and their standard errors of the yellowfin tuna the length frequency data from GFMM for the 1st quarter in 2014

parameters						
Age-groups	proportions		mean-lengths		standard deviations	
k	pi	(SE of pi)	mu	(SE of mu)	sigma	(SE of sigma)
1	0.03540	0.008058	45.29	0.3904	1.326	0.08668
2	0.13932	0.013742	49.92	0.2672	1.461	NA
3	0.35952	0.026362	54.64	0.2125	1.600	NA
4	0.45219	0.029243	58.59	0.1677	1.715	NA
5	0.01357	0.007442	64.04	1.0085	1.875	NA

Table 4.3: The complete set of standard errors for all the parameter estimates computed from the covariance matrix.

(SE of pi)	(SE of mu)	(SE of sigma)
0.008058	0.390354	0.086681
0.013742	0.267155	1.149126
0.026362	0.212482	1.261239
0.029243	0.167670	1.353019
0.007442	1.008475	1.482216

That is, the mixture model fails to reject the claim that yellowfin tuna population is indeed come from the GFMM distribution family with 5 distinct components/subpopulations.

Similar analyses were carried out for the remaining quarters and the Chi-square test of GOF and corresponding p-values reported in table 4.5. Examination of the p-values revealed that the fits were really good for the first two quarters, barely fit the the third quarter data and poorly fit the last quarter. This could be attributed to the fact that the correct number of subpopulations is five but not six as reported in the last two quarters and other factors to due missing data problem.

Based on the size selection used in this study, the proportions of successive age groups increase continuously until the fouth component. This may not be a typi-

Table 4.4: ANOVA and Chi-Squared test of goodness of fit for the parameter estimate for the 1st quarter

Analysis of Variance Table			
	Df	Chisq	Pr(>Chisq)
Residuals	3	4.726	0.193

Table 4.5: Chi-Square statistics with degrees of freedom, p-values, and the sample size for the four quarters

	Df	chisq	p-value	sample size
1st qter	3	4.726	0.19	1184
2nd qter	3	4.54	0.21	1185
3rd qter	7	19.741	0.01 **	1601
last qter	6	21.89	0.001 **	441

cal reflection of the entire population since in a stable population with mortality, the proportions of successive age groups are expected to be decreasing. Hence this conclusion could only arise from the size selection used in this fishery.

The chapter presented the findings and discussions resulting from the application of the GFMM to the 2014 yellowfin tuna fork length-frequency data. The summary from the histogram showed that the various components modes were heavily overlapped and hence the histogram alone is not reliable for estimating the number of components/cohorts.

The appropriate number of components and the initial values for all the parameters were guessed based on visual inspection of the histogram. Improved initial values were only chosen after the components densities were superimposed on the histogram and the sensitivities of the fits assessed graphically and objectively by the rootogram analysis and the Chi-square test of GOF from the ANOVA respectively. It was observed that the means and the standard deviations in particular were very sensitive when shifted in the horizontal and the vertical direction respectively.

The number of parameters to be estimated was quite high, therefore basic con-

straints were imposed on all the set of parameters. For instance, only four out of the five proportions were estimated by constraining the sum of the mixing proportions to one. Also, the CCV constriant imposed on the standard deviations enabled only the first standard error to be estimated and the remaining calculated from this one and all the estimated means by applying Delta method to the covariance matrix.

The analysis established that the ML estimation method when used in conjunction with the EM and NR algorithms is in general a powerful method for estimating age-groups related parameters if appropriate optimization constraints are imposed.

Chapter 5

CONCLUSION AND RECOMMENDATION

5.0 Introduction

This chapter presents the concluding statements from the study under review. It also provides some recommedations resulting from the findings to fisheries scientists and stakeholders in the fishery industry and suggestions on possible future research areas.

5.1 Conclusion

It was evident from the study under review that the yellowfin tuna population from the atlantic ocean of Ghana consists of five (or six) subpopulations with the same (normal or log normal) distribution for each component although different components distributions are theoretically viable. Hence, the GFMM significantly fits the yellowfin tuna length frequency data irrespective of the limitations of heavy overlapping of the components modes and the effect of different bins size. For each of the five or six component (age-group), all the parameters were successfully estimated by the ML method via the EM and the NR algorithms. First, ten

and twelve free parameters were estimated directly for the first two and last two quarters respectively. That is, for the first two quarters, four mixing proportions, five mean-lengths, and one standard deviation for each age-group were estimated directly. The remaining five (from the five subpopulations) and six (from the six subpopulations) were computed from the constraints imposed particularly on the proportions and the standarad deviations. Most of the estimates as revealed by their standered errors were statistically significant. Also, the Chi-square GOF test from the anova table revealed that the five components mixture model fit was statistically better than the six. Hence, based on the size selection in this fishery, it could be infered that there exist five different age groups with continuous increament in the mean lengths. The proportions and their equivalent estimated numbers also increased consecutively until the fourth component where it began to decrease.

Finally, the new variance covariance matrix (R-code) which was not originally part of the package together with the Chi-square GOF test addressed the inferencing problem encountered in the graphical methods. The new covariance matrix of dimension (15×15) estimated all the standard errors.

5.2 Recommendations

The following recommedations and suggestions are made to fisheries scientists, stakeholders and policy makers in the Fishery Commission:

- It is recommended that the GFMM and ML estimation method via EM and NR algorithms made available in the mixdist package in R should be chosen over the traditional Bhattacharya graphical method used by most fishery scientists in dissecting and estimating the age-groups related parameters from a grouped length frequency data.

- The Fishery Commission of Ghana should intensifies training and encourages its personel to take advantage of the flexibility and several packages in

R to augment its research work in fishery growth modeling, mortality rate and stock assessments models etc.

5.3 Suggestions on further research

- It is recommended that further investigation should be conducted into estimating the total number of the subpopulations (cohorts or components) from the mixture population since the scope of this work could not allow for that.

- Also, the effects of differents bins size or class widths on the estimates should be investigated

- Investigation into the feasibility of using only the the Newton Raphson algorithm (with quadratic convergence rate) as the required numerical method will be a big break through.

REFERENCES

Anandkumar, A., Hsu, D. and Kakade, S.M. (2012). A method of moments for mixture models and hidden markov models. *Work shop and Conference Proceedings vol* 23(2012)33.1 – 34

Baker, G.A. (1940). Maximum-likelihood estimation of the ratio of the two components of non-homogeneous populations. *Tohoku Math. J.*, 47, 304 – 308.

Beran, R. (1977). Minimum hellinger distance for parametric model. *Annals of statistics* 5, 445 – 463.

Bhattacharya, C. G. (1967). A simple method of resolution of a distribution into Gaussian components. *Biometrics*, 23 : 115 – 135.

Blight, B. J. N. (1970). Estimation from a censored sample for the exponential family. *Biometrika* 57, 389-395

Brady, C. M. (2008). Power analysis of finite mixture of poisson distributions. Thises

Buchanan-Wollaston, N J., and Hodgson, W. C. (1929). A new method of treating frequency curves in fishery statistics with some results. *J. Cons. Int. Explor. Mer.* 4 : 207 – 225.

Cassie, R. M. (1954). Some uses of probability paper in the analysis of size frequency distributions. *Aust. J. Mar. Freshwater Res.* 5 : 513 – 522.

Chen, Y. and Gupta, M. R. (2010), EM demystified: An expectation maximisation tutorial. *University of Washington, Dept. of EE, UWEETR*

Couvreur, C. (1995a). Estimation of parameters and classification of mixture of authoregressive processes. M.S. Thesis, University of Illinois at Urbana-Champaign

Couvreur, C. and Bresler, Y. (1995b). Decomposition of mixture of Gaussian AR process. Coordinated Science Laboratory and Department of Electrical and Computer Engineering University of Illinois at Urbana-Champaign 1308 West Main Street, Urbana, IL 61801, U.S.A

Day, N. E. (1969). Estimating the components of a mixture of normal distribution. *Biometrika* 56 (3) 463-474

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977), Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. B* 39, $1-38$.

Do, K. and McLachlan, G. J. (1984). Estimation of mixing proportions: a case study. *Applied Statistics* 33, 134–140.

Dong, Z. (1997). Mixture analysis and its preliminary application in Archaeology. *Journal of Archaeological Science* 24, $141-161$

Du, J. (2002). Combined algorithms for fitting finite mixture distributions. Master's thesis, Mc- Master University Hamilton, Ontaria.

Everitt, B.S., and Hand, D.J. (1981). Finite Mixture Distributions. Chapman and Hall, London.

Fowlkes, E.B. (1979). Some methods for studying the mixture of two normal (lognormal) distributions. *Journal of the American* Association 74, 561-73

Gayanilo, F.C. and Pauly, D. (1997). *The FAO − ICLARM Stock Assessment Tools(FiSAT) :Reference Manuel.* FAO, Rome.

Ghahramani, Z. and Jordan, M. I. (1994). Supervised learning from incomplete data via an EM approach. *Advances in Neutral Information Processing System. Morgan Kaufman Publishers, San Francisco, CA*

Magnifico, G. (2007). New insight into fish growth parameters estimation by means of Length-Based Method. *PhD theses. University of Rome.*

Haitovsky, Y. (1982). Grouped Data. *Encyclopedia of Statistical Sciences.* 3-527-536. Wiley, New York

Hald, A. (1952). Statisticasl theory with engineering applications $p.152-158\,Wiley$, $New\ York,\ N.Y.$ 783 $p.$

Harding, J. P.(1949). The use of probability paper for the graphical analysis of polymodal frequency distributions. *J. Mar. Biol. Assoc. U.K.*$28:141-153$

Harris, D. (1968). A method of separating two superimposed normal distributions using arithmetic probability paper. *J. Anim. Ecol.*, $37:\ 315-319.$

Hathaway, R. J. (1983). Constrained maximum-likelihood estimation for a mixture of m Univariate normal distributions. *Phd. Thesis. Rice University*

Hathaway, R. J. (1983). A constrained formulation of maximum likelihood estimation for normal mixture distributions. *The Annals of Statistics.* 13 (2) 795-800.

Heijan, D. F. (1989). Inference from Grouped Continuous Data. *A Review.Statistical Sciences.* 4 (2) 164-183

Holgersson, M. & Jorner, U. (1978). Decomposition of a mixture into normal components: a review. *International Journal of Bio−Medical Computing* 9, 367–392.

Jamshidian, M. and Jennrich, R. I. (1997). Acceleration of EM-algorithm by using Quasi-Newton methods. *Journal of the Royal Statisitical Society.* $59,\ 569-587$

Jamshidian, M. and Jennrich, R. I. (1993). Conjugate gradient acceleration of EM-algorithm. *Journal of the American Statistical Society.* $88,\ 221-228$

Kolding, J. and Ubal Giordano, W. (2002). Lecture notes. Report of the AdriaMed training course on fish population dynamics and stock assessment.

FAO MiPAF Scientific Cooperation to Support Responsible Fisheries in the Adriatic Sea. GCP/RER/010/ITA/TD-08. *AdriaMed Technical Documents*, 8 : 143 *p.*

Laslett, G. M., Eveson, J. P., and Polacheck, T. (2004). Fitting growth models to length frequency data. *ICES Journal of Marine Science*, 61: 218-230.

Louis, T. A. (1982). Finding the Observed Information Matrix when Using EM Algorithm.*Joun. of Royal Statistical Society. Series B.* 44, 226-233.

MacDonald, P. and Green, P. (1988). User's Guide to Program MIX: An Interactive Program for Fitting Mixtures of Distributions, release 2.3. ICHTHUS DATA SYSTEMS, Ontario, Canada

MacDonald, P.D.M. and Pitcher, T.J. (1979). Age-groups from size-frequency data: a versatile and efficient method of analyzing distribution mixtures. *J. Fish. Res. Board Can.*, 36 : 987 − 1001

MacDonald, P.D.M. and Du, J. (2011). Finite mixture distribution models. Package "*mixdist*".

Macdonald, P.D.M. and Green, P.E.J. (1988). *User's Guide to Program MIX : An Interactive Program for Fitting Mixtures of Distributions*, *Release*2.3. ICHTHUS DATA SYSTEMS

McLachlan, G. and Basford, K. E. (1988). Mixture Models: Inference and Applications to Clustering. Dekker, New York.

McLachlan, G. and Peel, D. (2000a). Finite Mixture Models. New York: Wiley.

Oka, M. (1954). Ecological studies on the Kidai by the statistical method. II. On the growth of the Kidai| (Taius tumifrons). *Bull. Fac. Fish. Nag* − *asakl* 2 : 8 − 25

Meng, X. L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* 80, 2, 267 − 278

Mendenhall, W. and Hader, R.J. (1958) Estimation of parameters of mixed ex-

ponentially distributed failure time distributions from censored life test data. *Biometrika* 45, 504 – 520.

Neyman, J. (1949). Contribution to the theory of χ^2 test. *Proc. (First) Berkeley symp. on Math. Statist. Prob.*, 239 – 279

Nylund, K. L., Asparouhov, T. and Muthén, B. O. (2007). Deciding on the number of classes in latent analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling A Multidisciplinary Journal*. 14 (4), 535–569

Pauly, D. and J.F. Caddy, 1985. A modification of Bhattacharya's method for the analysis of mixtures of normal distributions. *FAO Fish. Circ.*, 781 : 781: 16 p. Pauly, D. and Morgan, G.R. (1987). In Length-Based Methods in Fisheries Research (Editors) *ICLARM Conference Proceedings*, *Kuwait*

Pearson, K. (1894). Contribution to the theory of mathematical evolution. *Philosophical Transaction of the Royal Society of London A* 185, 71 – 110.

Petersen, C.G.J. (1891). Eine Methode zur Bestimmung des Alters und des Wuchsesder Fisches. *Mitt .Dtsch. Seefisch. Ver.*, 11 : 226 – 235

Quandt, R. E. and J. B. Ramsey (1978). Estimating Mixtures of Normal Distribu tions and Switching Regressions. *Journal of the American Statistical Association*, 73, 730 – 738.

Redner, R.A., and Walker, H.F. (1984). Mixture densities, maximum likelihood, and the EM algorithm. *Society for Industrial and Applied Mathematics Review* 26, 195 – 239.

Rao, C.R. (1948). The utilization of multiple measurements in problems of biological classification. *J. Roy. Statist. Soc. B*, 10 : 159 – 203.

Sundberg, R. (1996). Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics*. 1 : 49 – 58

Sundberg, R. (1976). An iterative method for solution of the likelihood equations

for incomplete data from exponential family. *Communication in Statistics-Simulation and Computation*. $B5$ (1), $55-64$

Tanaka, S. (1962). A method of analysing a polymodal frequency distribution and its application to the length distribution of the porgy, Taius tumif rons (T. and S.). *J. Fish. Res. Board Can.* $19: 1143-1159$.

Taylar, B. J. R. (1965). The analysis of polymodal frequency distributions. *J. Anim. Ecol.* $34: 445-452$.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). Statistical analysis of finite mixture distributions. John Wiley & Sons Ltd, Chichester.

Wang, H. x., Luo, B., Zhang, Q. b., and Wei, S. (2004). Estimation for the number of components in a mixture model using stepwise split-and-merge EM algorithm. *Pattern Recognition Letters*. 25, $1799-1809$

Wirjanto,T. S. and Xu, D. (2009). The application of mixture of normal distri butions in emperical finance: A selected survey, University of Waterloo.

Woodward, W. A. and Eslinger, P. W. (1983). Minimum Hellinger distance estimation of mixture of model parameter. *NASA Technical Report. SR*-63 -04433

Woodbury, M. A. (1971). Discussio of paper by Hartley and Hocking. *Biometrika* 27, $808-817$

Woodward, W.A., Parr, W.C., Schucany, W.R., and Lindsey, H. (1984) A com parison of minimum distance and maximum likelihood estimation of a mix ture proportion.*Journal of the American Statistical Association* 79, 590-598

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics* 11, $95-103$

Xu, L. and Jordan M. I. (1995). On the convergence properties of the EM Algorithm for Guassian mixture. *A. I. Memo.* 1520

Ynrowtrz, S. J. (1969). A consistent estimator for the identification of finite models. *Ann. Math. Stat.* 40: 1728-1735

Zhu, X. and Zhao, Y. (2013). Length Frequency Age Estimations of American Eel Recruiting to the Upper St. Lawrence River and Lake Ontario. *The Transaction of the American Fisheries Society*. 142: 333-344

Appendices

Appendix A

First Quarter Yellowfin tuna Fork Length frequency data.

Fork Lengths in (cm)	Frequencies	Species	Gear-type	School-type
43	3	YFT	purse seine	free school
44	5	YFT	purse seine	free school
45	9	YFT	purse seine	free school
46	12	YFT	purse seine	free school
47	13	YFT	purse seine	free school
48	16	YFT	purse seine	free school
49	28	YFT	purse seine	free school
50	41	YFT	purse seine	free school
51	43	YFT	purse seine	free school
52	46	YFT	purse seine	free school
53	57	YFT	purse seine	free school
54	79	YFT	purse seine	free school
55	99	YFT	purse seine	free school
56	136	YFT	purse seine	free school
57	122	YFT	purse seine	free school
58	106	YFT	purse seine	free school

Source: Fishery Commission of Ghana (Marine Section, Tema, 2014)

First Quarter Yellowfin tuna Fork Length frequency data.

Fork Lengths in (cm)	Frequencies	Species	Gear-type	School-type
59	140	YFT	purse seine	free school
60	109	YFT	purse seine	free school
61	54	YFT	purse seine	free school
62	38	YFT	purse seine	free school
63	13	YFT	purse seine	free school
64	7	YFT	purse seine	free school
65	4	YFT	purse seine	free school
66	2	YFT	purse seine	free school
67	1	YFT	purse seine	free school
68	1	YFT	purse seine	free school
69	1	YFT	purse seine	free school
70	0	YFT	purse seine	free school

Source: Fishery Commission of Ghana (Marine Section, Tema, 2014)

Second Quarter Yellowfin tuna Fork Length frequency data.

Fork Lengths in (cm)	Frequencies	Species	Gear-type	School-type
37	0	YFT	purse seine	free school
38	1	YFT	purse seine	free school
39	1	YFT	purse seine	free school
40	1	YFT	purse seine	free school
41	1	YFT	purse seine	free school
42	2	YFT	purse seine	free school
43	2	YFT	purse seine	free school
44	2	YFT	purse seine	free school
45	2	YFT	purse seine	free school
46	3	YFT	purse seine	free school
47	5	YFT	purse seine	free school
48	9	YFT	purse seine	free school
49	19	YFT	purse seine	free school
50	36	YFT	purse seine	free school
51	44	YFT	purse seine	free school
52	54	YFT	purse seine	free school
53	89	YFT	purse seine	free school
54	138	YFT	purse seine	free school
55	153	YFT	purse seine	free school
56	230	YFT	purse seine	free school
57	205	YFT	purse seine	free school
58	186	YFT	purse seine	free school
59	174	YFT	purse seine	free school
60	147	YFT	purse seine	free school
61	52	YFT	purse seine	free school

Source: Fishery Commission of Ghana (Marine Section, Tema, 2014)

Second Quarter Yellowfin tuna Fork Length frequency data.

Fork Lengths in (cm)	Frequencies	Species	Gear-type	School-type
62	24	YFT	purse seine	free school
63	10	YFT	purse seine	free school
64	4	YFT	purse seine	free school
65	2	YFT	purse seine	free school
66	0	YFT	purse seine	free school
67	1	YFT	purse seine	free school
68	1	YFT	purse seine	free school
69	2	YFT	purse seine	free school
70	0	YFT	purse seine	free school
71	1	YFT	purse seine	free school
72	1	YFT	purse seine	free school
73	1	YFT	purse seine	free school
74	1	YFT	purse seine	free school
75	1	YFT	purse seine	free school

Source: Fishery Commission of Ghana (Marine Section, Tema, 2014)

Third Quarter Yellowfin tuna Fork Length frequency data.

Fork Lengths in (cm)	Frequencies	Species	Gear-type	School-type
39	0	YFT	purse seine	free school
40	1	YFT	purse seine	free school
41	1	YFT	purse seine	free school
42	1	YFT	purse seine	free school
43	2	YFT	purse seine	free school
44	3	YFT	purse seine	free school
45	7	YFT	purse seine	free school
46	12	YFT	purse seine	free school
47	17	YFT	purse seine	free school
48	24	YFT	purse seine	free school
49	82	YFT	purse seine	free school
50	61	YFT	purse seine	free school
51	35	YFT	purse seine	free school
52	18	YFT	purse seine	free school
53	7	YFT	purse seine	free school
54	4	YFT	purse seine	free school
55	1	YFT	purse seine	free school
56	1	YFT	purse seine	free school
57	0	YFT	purse seine	free school
58	1	YFT	purse seine	free school
59	2	YFT	purse seine	free school
60	2	YFT	purse seine	free school
61	4	YFT	purse seine	free school
62	5	YFT	purse seine	free school
63	7	YFT	purse seine	free school
64	12	YFT	purse seine	free school

Source: Fishery Commission of Ghana (Marine Section, Tema, 2014)

Third Quarter Yellowfin tuna Fork Length frequency data.

Fork Lengths in (cm)	Frequencies	Species	Gear-type	School-type
65	18	YFT	purse seine	free school
66	26	YFT	purse seine	free school
67	18	YFT	purse seine	free school
68	13	YFT	purse seine	free school
69	15	YFT	purse seine	free school
70	6	YFT	purse seine	free school
71	9	YFT	purse seine	free school
72	5	YFT	purse seine	free school
73	4	YFT	purse seine	free school
74	7	YFT	purse seine	free school
75	5	YFT	purse seine	free school
76	16	YFT	purse seine	free school

Source: Fishery Commission of Ghana (Marine Section, Tema, 2014)

Appendix B

R Codes

Install and load "mixdist" package from R cran repository

```
> install.packages("mixdist")
```

```
> library(mixdist)
```

Setting the working directory

```
> setwd("C://Users/Peter/Desktop")
```

Loading the data

```
> fqtuna = read.csv("fquotadata.csv", header=T, na.strings="")
```

```
> fix(fqtuna)
```

Grouping the data into class width of size $h = 2$

```
> fqtunagrp = mixgroup(fqtuna, breaks=c(0, seq(44,70,2), 75))
```

alternatively, loading the already grouped data into R

```
> fqtunagrp = data.frame(length=c(44,46,48,50,52,54,56,58,
  + 60,62,64,66,68,Inf),freq=c(8,21,29,69,89,136,235,228
  ,249,92,20,6,1,1))
```

```
> class(fqtunagrp)= c("mixdata", "data.frame")
```

a function for plotting grouped length frequency histogram for `fqtunagrp`

```
> plot.mixdata(fqtunagrp, xlab="Fork Length (cm)",ylab="Relative
     frequency", main="First Quarter (2014)")
```

Choosing the initial values with the function mixparam(mu,sigma, pi)

```
> fqtunapik=param( mu=c(45,50,55,60,65),
     + sigma = c(1.3,1.4,1.5,1.6,1.7), pi=rep(0.4,5))
```

fitting the distribution on the histogram with the initial values

```
> fitfqtunagrp= mix(fqtunagrp ,dist="norm"
     ,constr=mixconstr(consigma="CCV"),emsteps=20)
```

Rootogram analysis

```
> plot(fitfqtunagrp, main="1st quarter(2014)", root=T)
```

Chi-square goodness of fit test through the anova function

```
> anova(fitfqtunagrp)
```

display of parameter values, standard errors, distribution, ANOVA

```
> summary(fitfqtunagrp)
```

Variance Covariance matrix and their corresponding standard deviation for the 10 free parameters estimated

```
> vmat.ccv = fitfqtunagrp$vmat
> sqrt(diag(vmat.ccv))
```

function that compute the full covariance matrix from the Delta method and the constriant CCV which is not part of the original mixdist package

```
> vmat.cccv = function(fitfqtunagrp) {
     k = nrow(fitfqtunagrp$parameters)
     sigma = fitfqtunagrp$parameters$mu
     Dmat = matrix(0, nrow = 3 * k, ncol = 2 * k)
     for (i in 1:(k - 1)) {
        Dmat[i, i] = 1
        Dmat[k, i] = -1
     }
     Dmat[k + 1, k] = 1
     Dmat[(2 * k + 2):(3 * k), k] =-sigma[1] * mu[-1]/mu[1] * 2
```

```
      for (i in (k + 1):(2 * k - 1)) {
         Dmat[i + 1, i] = 1
         Dmat[i + k + 1, i] = sigma[1]/mu[1]
      }
      Dmat[(2 * k + 1):(3 * k), 2 * k] = mu/mu[1]
      Dmat %*% fitfqtunagrp$vmat %*% t(Dmat)
}
vmat.cccv
> vmat.seq
function(fitfqtunagrp) {
      k = nrow(fitfqtunagrp$parameters)
      Dmat = matrix(0, nrow = 3 * k, ncol = 2 * k)
      for (i in 1:(k - 1)) {
         Dmat[i, i] = 1
         Dmat[k, i] = -1
      }
      for (i in k:(2 * k - 1)) {
         Dmat[i + 1, i] = 1
      }
      Dmat[(2 * k + 1):(3 * k), 2 * k] = 1
      Dmat %*% fitfqtunagrp$vmat %*% t(Dmat)
}
> vmat.cccv(fitfqtunagrp)
> sqrt(diag(vmat.cccv(fitfqtunagrp)))
```

Appendix C

Variance Covariance Matrix