ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS

ESTIMATION AND TESTING PROBLEMS IN POISSON MIXTURES

By

Dimitris Karlis

A THESIS

Submitted to the Department of Statistics

of the Athens University of Economics and Business

in partial fulfillment of the requirements for

the degree of PhD in Statistics

Athens, Greece

August 1998

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΠΡΟΒΛΗΜΑΤΑ ΕΚΤΙΜΗΣΗΣ ΚΑΙ ΕΛΕΓΧΩΝ ΥΠΟΘΕΣΕΩΝ ΣΕ ΠΕΠΕΡΑΣΜΕΝΑ ΜΕΙΓΜΑΤΑ ΤΗΣ ΚΑΤΑΝΟΜΗΣ POISSON

Δημήτρης Καρλής

ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής

του Οικονομικού Πανεπιστημίου Αθηνών

ως μέρος των απαιτήσεων για την απόκτηση

Διδακτορικού Διπλώματος στη Στατιστική

Αθήνα
Αύγουστος 1998

to my father

Acknowledgements

I never felt alone during the period of this research. Many people offered me their help in several ways. My family supported me with patience during this thesis. I feel indebted to Professor Evdokia Xekalaki, my supervisor, for continuous encouragement and helpful comments that helped to improve substantially the contents of this thesis.

I would also thank the National Scholarship Foundation for financial support in the framework of the 29th program of internal scholarships.

Special thanks to the staff of the Department of Statistics, Athens University of Economics and Business, especially for their support in difficult personal circumstances. I also really appreciated the help of my colleagues in the laboratory of the Department of Statistics.

I would also like to pay tribute to my unforgettable friend Nikos Louganis who, unfortunately, passed away the summer of '96 for many interesting discussions about Statistics.

Finally, I would like to thank Athina Karvounaraki, Nikos Diamadidis and Dimitris Koutroulis for their friendship. Many thanks to all who provided me with papers not available in our library (and they are too many).

I am sure that beyond this thesis, it is more important for me living, discussing and communicating with all the above mentioned persons. I acknowledge that the route to this thesis was the most interesting and edifying experience, more interesting than the thesis itself.

Abstract

Mixture models is a rapidly developing area of statistics with applications to a variety of fields. This thesis is devoted to Poisson mixtures which naturally arise as alternative models when the simple Poisson model fails to describe the data. For example, it is known that the Poisson distribution is characterised by its property of having a variance equal to its mean. This property is not usually satisfied by the data. This case is usually referred to as overdispersion. Poisson mixtures can provide flexible alternative models that can represent the inhomogeneity of the population. The idea is that the persons comprising the entire population do not have the same Poisson parameter. Instead, their parameter varies according to a distribution, termed as the mixing distribution. By the law of total probability, Poisson mixtures arise. The properties of Poisson mixtures are examined in depth.. Due to their complexity, only a few have been examined in the literature. Several members of this family are presented in this thesis, emphasising their interrelations.

Among these models finite Poisson mixtures are very popular since they admit a simple and natural interpretation, as models describing a population consisting of a finite number of subpopulations. Moreover, even if the true mixing distribution is continuous, one is restricted to estimate it via a finite distribution. Estimation methods for finite Poisson mixtures are explored. Two distinct cases appear in practice. The first assumes that the number of components is fixed and tries to estimate the parameters maximising a criterion over the space of all mixing distributions with the given number of support points. The second, termed as the semiparametric case, treats the number of components as an unknown parameter which has to be estimated from the data. For the first case, the EM algorithm for maximum likelihood estimation is an inexpensive and easy method with numerous applications. In this thesis the algorithm is critically reviewed. Initial values that can help the algorithm to converge quicker are examined via a simulation experiment. An improvement of the EM algorithm for mixtures based on a property of mixtures from the exponential family is proposed. For the semiparametric case, the algorithms proposed for obtaining maximum likelihood estimates are examined. These

algorithms do not seem to be adequate for the case of Poisson mixtures since the number of support points is usually small, and the algorithms do not work properly.

The problem of determining the number of components is also examined. A new method is proposed. The method is based on sequentially applying the likelihood ratio test using bootstrap methods to determine the distribution of the test statistic. The properties of this new method are also examined.

Several other methods of estimation are also reviewed. For the moment method, the existence of the estimates is explored. The results show that the moment estimates do not exist very often. Moreover, the small sample comparison of the moment estimators to the maximum likelihood estimators discourage their use. An alternative method which uses the zero frequency instead of the third moment is developed. This method is useful when the zero proportion is large.

A new method, which is efficient and robust at the same time is introduced. The method is based on minimising the Hellinger distance. The obtained estimators are examined and shown to be robust relative to the maximum likelihood estimators. This robustness property is used for proposing inferential procedures for Poisson mixtures. It is proposed to use the Minimum Hellinger Estimators for semiparametric estimation. Moreover, diagnostic graphs can be used for detecting if the Poisson distribution is appropriate. These graphs are not influenced by a few observations and can, thus, detect if a Poisson distribution is appropriate. In addition, an alternative to the likelihood ratio test is proposed. This is termed as the Hellinger Deviance Test and is based on the difference of the Hellinger distance between two hypotheses. This test statistic is powerful and robust to outlier contamination. An algorithm for the estimation of the parameters is provided which facilitates the application of minimum Hellinger methodologies

ΠΕΡΙΛΗΨΗ

Η σημασία των μοντέλων μειγμάτων κατανομών στη Στατιστική είναι μεγάλη και το πεδίο εφαρμογών τους συνεχώς αυξάνεται. Η διατριβή αυτή ασχολείται με μείγματα της κατανομής Poisson τα οποία χρησιμοποιούνται ως εναλλακτικά μοντέλα στις περιπτώσεις που η απλή κατανομή Poisson αποτυγχάνει να περιγράψει τα δεδομένα.. Για παράδειγμα, είναι γνωστό ότι η κατανομή Poisson έχει τη χαρακτηριστική ιδιότητα ότι η διακύμανση της είναι ίση με την αναμενόμενη τιμή της. Αυτό πολλές φορές δεν συμβαίνει στην πράξη. Τα μείγματα της κατανομής Poisson αποτελούν ευέλικτα εναλλακτικά μοντέλα που μπορούν να περιγράψουν την ανομοιογένεια του πληθυσμού. Η λογική τους στηρίζεται στο γεγονός ότι εξαιτίας της ανομοιογένειας του πληθυσμού, τα άτομα που συνιστούν τον πληθυσμό δεν έχουν την ίδια συχνότητα εμφάνισης του υπο εξέταση γεγονότος. Η συχνότητα αυτή περιγράφεται από την παράμετρο της κατανομής Poisson, και συνεπώς ο καθένας μπορεί να έχει μια διαφορετική τιμή για τη συχνότητα αυτή. Επομένως η συχνότητα εμφάνισης του γεγονότος που περιγράφεται από την παράμετρο της κατανομής Poisson είναι μια τυχαία μεταβλητή που ακολουθεί κάποια κατανομή, η οποία αποκαλείται κατανομή μίξης. Τότε από το θεώρημα ολικής πιθανότητας τα μείγματα της κατανομής Poisson προκύπτουν. Οι ιδιότητες τους εξετάζονται σε αυτή τη διατριβή. Λόγω της πολυπλοκότητας τους, μόνο λίγες τέτοιες κατανομές έχουν ερευνηθεί. Σε αυτή τη διατριβή ένας μεγάλος αριθμός μελών της οικογένειας αυτής παρουσιάζεται και δίνεται έμφαση στις μεταξύ τους σχέσεις.

Μεταξύ των μειγμάτων της κατανομής Poisson τα πεπερασμένα μείγματα αποτελούν μια ενδιαφέρουσα κατηγορία. Είναι ιδιαίτερα διαδεδομένα κυρίως λόγω της απλής φυσικής ερμηνείας τους ως μοντέλα που περιγράφουν έναν πληθυσμό με πεπερασμένο αριθμό υποπληθυσμών. Επιπλέον, ακόμα και στις περιπτώσεις στις οποίες η κατανομή μίξης είναι συνεχής, η εκτίμησή της συνίσταται στην εκτίμηση μιας κατανομής με πεπερασμένο αριθμό σημείων με μη αρνητική πιθανότητα. Στη διατριβή αυτή εξετάζονται μέθοδοι εκτίμησης για πεπερασμένα μείγματα της κατανομής Poisson. Στην πράξη εμφανίζονται δυο διαφορετικές περιπτώσεις. Στην πρώτη ο αριθμός των μελών του μείγματος είναι δεδομένος οπότε απαιτείται η βελτιστοποίηση ενός κατάλληλου κριτηρίου για όλες τις κατανομές μίξης με το δεδομένο αριθμό σημείων με μη αρνητική πιθανότητα. Στη δεύτερη περίπτωση, ο

αριθμός των μελών του μείγματος είναι άγνωστος και πρέπει επομένως να εκτιμηθεί από τα δεδομένα. Η περίπτωση αυτή είναι γνωστή ως ημιπαραμετρική περίπτωση. Για την πρώτη περίπτωση, ο αλγόριθμος ΕΜ προσφέρεται για εκτίμηση με τη μέθοδο μεγίστης πιθανοφάνειας και έχει χρησιμοποιηθεί ευρέως σε πολλές εφαρμογές. Στη διατριβή αυτή γίνεται μια κριτική επισκόπηση των χρήσεων του αλγορίθμου. Το πρόβλημα της επιλογής αρχικών τιμών που μπορούν να βελτιώσουν την ταχύτητα του αλγορίθμου εξετάζεται επίσης μέσω προσομοίωσης. Επιπλέον χρησιμοποιώντας ιδιότητες των μειγμάτων κατανομών της εκθετικής οικογένειας κατανομών προτείνεται μια νέα μέθοδος που βελτιώνει την ταχύτητα του αλγορίθμου. Στην ημιπαραμετρική περίπτωση παρουσιάζονται οι αλγόριθμοι που υπάρχουν στη βιβλιογραφία. Οι αλγόριθμοι αυτοί δεν είναι κατάλληλοι για την περίπτωση των μειγμάτων Poisson γιατί στα μείγματα αυτά ο αριθμός των μελών του μείγματος είναι συνήθως μικρός με αποτέλεσμα να αποτυγχάνουν οι αλγόριθμοι.

Το πρόβλημα της εκτίμησης του αριθμού των μελών του μείγματος εξετάζεται επίσης. Μια νέα μέθοδος προτείνεται η οποία βασίζεται στην διαδοχική χρήση του ελέγχου λόγου πιθανοφανειών με μεθόδους bootstrap. Η χρήση των μεθόδων bootstrap είναι απαραίτητη γιατί η κατανομή της ελεγχοσυνάρτησης είναι άγνωστη. Οι ιδιότητες της μεθόδου ερευνώνται και προσδιορίζονται συνθήκες που διευκολύνουν τη χρήση του ελέγχου.

Εναλλακτικές μέθοδοι εκτίμησης ερευνώνται επίσης. Για τη μέθοδο των ροπών αποδεικνύεται ότι πολύ συχνά οι εκτιμήτριες δεν υπάρχουν επειδή το σύστημα των εξισώσεων δεν έχει λύση. Με βάση τις συγκρίσεις με τη μέθοδο της μεγίστης πιθανοφάνειας για μικρά δείγματα και αποδεικνύεται ότι η μέθοδος των ροπών έχει χαμηλότερη απόδοση. Μια εναλλακτική μέθοδος που εξετάζεται χρησιμοποιεί την παρατηρούμενη συχνότητα της τιμής 0 αντί της τρίτης ροπής για τις περιπτώσεις όπου η παρατηρούμενη σχετική συχνότητα της τιμής 0 είναι σχετικά μεγάλη.

Μια νέα μέθοδος, η οποία είναι ταυτόχρονα αποτελεσματική και εύρωστη μελετήθηκε. Η μέθοδος βασίζεται στην ελαχιστοποίηση της απόσταση Hellinger. Οι εκτιμήτριες που λαμβάνονται εξετάζονται και αποδεικνύεται ότι είναι εύρωστες σε σχέση με τις εκτιμήτριες μεγίστης πιθανοφάνειας. Η ιδιότητα αυτή μπορεί να χρησιμοποιηθεί για την ανάπτυξη μεθοδολογιών στατιστικής συμπερασματολογίας βασισμένων στην απόσταση Hellinger, όπως ημιπαραμετρική εκτίμηση, διαγνωστικά γραφήματα τα οποία έχουν σκοπό να διαγνώσουν αν η απλή κατανομή Poisson είναι

κατάλληλη και ελέγχους υποθέσεων εναλλακτικούς των έλεγχοι υποθέσεων βασισμένων στο λόγο πιθανοφανειων. Τέτοιοι έλεγχοι έχουν μεγάλη ισχύ και είναι ανθεκτικοί στην παρουσία ακραίων τιμών

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

General Introduction

1.1 An Introduction to Mixture Models

Mixture models are widely used in statistical modelling since they can model situations which a simple model cannot describe. For example, assuming a specific distribution $F(\cdot|\theta)$ for a data set means that the mean to variance relation is given for this distribution. In practical situations this may not be true. A simple example is the Poisson distribution. It is well known (see, e.g. Johnson *et al.*, 1992) that for the Poisson distribution the variance is equal to the mean and this property characterises the Poisson distribution among all the discrete distributions. Hence assuming a Poisson distribution is equivalent to assuming a distribution with the mean equal to its variance. With real data sets this may not be true. Often the sample mean is noticeably exceeded by the sample variance. This situation is known as overdispersion. Evidently, a Poisson distribution will not be a suitable model. The need of a more general family of distributions is obvious in these cases. Such a flexible family may arise if we consider the parameter (or the parameters) θ of the original distribution as varying according to a distribution with probability density function say $g(\theta)$.

Definition 1.1 A distribution function $F(\cdot)$ is called a mixture of the distribution function $F(\cdot|\theta)$ with mixing distribution $G(\theta)$ if it can be written in the form

$$F_x(x) = \int_{\Theta} F_{x|\theta}(x|\theta) dG_\theta(\theta)$$

where Θ is the space in which è takes values. The above definition can be also expressed in terms of probability density functions (or the probability functions in the discrete case), thus

$$f_x(x) = \int_{\Theta} f_{x|\theta}(x) g_\theta(\theta) d\theta \tag{1.1}$$

We will denote the mixture in (1.1) as $f(x|\theta) \underset{\theta}{\Lambda} g(\theta)$. The density $g_\theta(\theta)$ is referred to as the mixing density.

Note that the mixing distribution is not necessarily continuous. It may be discrete or even a distribution with positive probability at finite points, i.e. a finite step distribution. In the sequel a mixture with a finite step mixing distribution will be termed a k-finite step mixture of $F(\cdot|\theta)$, where k is a non-negative integer referring to the number of points with positive probabilities in the mixing distribution.

A physical interpretation of mixture models is that the i-th individual of the population has a distribution defined by a probability density function $f(x|\theta_i)$. All the members of the population follow the same parametric form of distribution but the parameter $è_i$ varies from individual to individual according to a distribution $G(\theta)$. Depending on the choice of the mixing distribution $G(\theta)$, an extremely broad family of distributions is obtained, which may be adequate for cases where the simple model fails. So, a mixture model describes an inhomogeneous population while the mixing distribution describes the inhomogeneity of the population. If the population was homogeneous, then all the members would have the same parameter θ, and the simple model would adequately describe the situation.

Mixture models cover several distinct fields of the statistical science. Their broad acceptance as adequate models to describe diverse situations, is evident from the plethora of their applications in the statistical literature. Titterington *et al.* (1985) gave a long list of papers with applications of mixture models up to 1985. In recent years the number of applications increased mainly because of the availability of high speed computer resources which removed any obstacles to apply such methods. Thus, mixture models have found applications in as diverse fields as:

- ***Data modelling***. With a data set in hand, usually, we need to fit a distribution, which can provide information concerning the underlying mechanism. As mixture

models are widely used to describe the inhomogeneity of a population they have became a very popular choice in practice. As already mentioned, mixture models are often called overdispersion models, because of their suitability to describe overdispersed data. The derivation of the negative binomial distribution, as a mixture of the Poisson distribution with a gamma distribution as the mixing distribution, originally given in Greenwood and Yule (1920) constitutes a typical example. Generally speaking mixture models can serve as an unrestricted platform for describing datasets for which simpler models fail to account.

- ***Discriminant analysis***. In discriminant analysis, the need of a method which can discriminate the population from which a new observation comes is obvious. Assuming a finite mixture model we may obtain the parameters of the subpopulations from a training set, and then classify the new observations via simple probabilistic arguments (see, e.g. McLachlan, 1992).
- ***Cluster analysis***. Cluster analysis is another interesting field where mixture models find applications. In McLachlan and Basford (1988) the reader can find a full description of the use of mixture models in cluster analysis. The idea is to describe the entire population as a mixture model consisting of several subpopulations (clusters). Then, by fitting a mixture model we may obtain the posterior probability for each observation to belong to any of the subpopulations (clusters).
- ***Outlier-robustness studies***. Outliers in data sets have been modelled with mixture models (see, e.g. Aitkin and Wilson, 1980). It is assumed that outliers comprise a component in a mixture model. Hence, by fitting a mixture model we may investigate the existence of outliers. In robustness studies, the contamination of the data can also be regarded as an additional component of a mixture model.
- ***ANOVA.*** The well known technique of the analysis of variance is a particular application of mixture models. The model assumes that the mean of the normal distribution of the entire population, varies from subpopulation to subpopulation. Then, it is assumed that the mean is itself a normal variate and the decomposition of the total variance is with respect to randomness and mixing. In general, all the mixed effect models are mixture models.
- ***Kernel density estimation***. In kernel density estimation the aim is to estimate the density of a sample by imposing n kernels $K(x)$ on the observations. Usually,

$K(x)$ is itself a symmetric probability density function and each kernel is given a weight equal to 1/n (see, e.g. Silverman, 1986). Thus, in kernel density estimation a mixture model is considered for the kernel with equal mixing probabilities. More specifically, the density estimate at point x is calculated as

$$f_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{1.2}$$

where h is a switching parameter which handles the smoothing procedure. The above form clearly represents a n-finite mixture of the kernel $K(x)$.

- ***Latent structure models.*** In latent structure models it is assumed that beyond the observable random variables there are other unobservable or even unmeasurable variables, which influence the situation under investigation. This has a common element with the method of factor analysis for continuous variables. The main assumption in the case of latent structure models is that of conditional independence, i.e. the assumption that for a given value of the unobservable variable the remaining variables are independent. Since inference is based on the unconditional distribution, we obtain, by the law of total probability, a mixture model where the mixing distribution represents the distribution of the unobservable quantity which thus is of special interest in many situations (see, e.g. Everitt, 1984a). It is very interesting that many methods proposed for mixture models are applicable to latent variable models (see, e.g. Aitkin *et al.*, 1981).
- ***Empirical Bayes Estimation.*** In Bayesian estimation the parameters of any distribution are considered to be random variates having their own distribution, known as the prior distribution. The prior distribution corresponds to the mixing distribution in (1.1). Specifically, the empirical Bayesian methods aim at estimating the prior distribution from the data. This obvious relation between these two distinct areas of statistics have resulted in a vast number of papers in both areas, with many common elements.
- ***Bayesian statistics***. Mixtures have been proposed to be used as priors, the main reason being their flexibility (see, e.g. Dalal and Hall, 1983). Beyond that, such priors are also robust and have been proposed for examining Bayesian robustness (Bose, 1994).

- ***Random variate generation***. The mixture representation of some distributions is a powerful tool for efficient random number generation from these distributions. Several distributions (discrete or continuous) may arise as mixture models from certain distributions, which are easier to generate. Hence generating variables from such representation can be less expensive. For more details, the reader is referred to Devroye (1992).
- ***Approximation of the distribution of some statistic***. In many statistical methods, the derived statistics do not have a standard distributional form and an approximation have to be used for their distribution. Mixture models allow for flexible approximation in such cases. Such an example is the approximation of the distribution of the correlation coefficient used in Gupta and Huang (1981).

Interesting reviews for the mixture models are given in the books by Everitt and Hand (1981), Titterington *et al.* (1985), McLachlan and Basford (1988), Lindsay (1995) as well as in the review papers of Gupta and Huang (1981), Redner and Walker (1984) and Titterington (1990).

1.2 Some Properties of Mixture Models

Mixture models have interesting properties. In this section some of their properties that will be used in the sequel are briefly discussed. Some definitions, notation and terminology are also provided.

Associative Property

It can easily be shown that

$$\left[f(x|\theta) \underset{\theta}{\Lambda} g(\theta|\mu) \right] \underset{\mu}{\Lambda} h(\mu) \quad \text{is equivalent to} \quad f(x|\theta) \underset{\theta}{\Lambda} \left[g(\theta|\mu) \underset{\mu}{\Lambda} h(\mu) \right]$$

The proof is based on the definition of such mixtures and the possibility of changing the order of integration or summation. In order that the above formula be valid we need to assume that there are no dependencies between the parameters of the

distributions considered. For example, $h(\mu)$ does not depend on θ. Then, the associative property holds and the order of mixing can be changed.

Moments of Mixture Distributions

Regardless of the form of $f(x|\theta)$ the expected value of the function $w(x)$ of the mixed distribution is obtained as:

$$E[w(X)] = \int_{\Theta} E_{x|\theta}[w(X)]g(\theta)d\theta \tag{1.3}$$

with the subscript in the expectation denotes that the expectation is taken with respect the conditional distribution of X. The integration must be replaced by summation in the case of discrete mixing distribution. It follows that

$$E[X] = E\left[E_{x|\theta}(X)\right]$$

$$V[X] = V\left[E_{x|\theta}(X)\right] + E\left[V_{x|\theta}(X)\right] \tag{1.4}$$

i.e. the variance of the mixed variate is the sum of the variance of the conditional mean plus the mean of the conditional variance. Relationship (1.4) shows that the mixture model has always a larger variance than the simple model and this explains the use of the term overdispersion models used for mixture models.

Mixture Models and Products of Random Variates

Another interesting property has been given by Sibuya (1979).

Proposition 1.1 (Sibuya, 1979). Mixing with respect to a scale parameter is equivalent to obtaining the distribution of the product of two random variables; the distribution function of the first variable is the same distribution as the conditional but with unit scale parameter, while the distribution of the second variable is the mixing distribution.

The above proposition justifies the derivation of certain distributions using both approaches; as mixtures and as the product of two random variables.

Some definitions of certain concepts used in the sequel are now provided.

Definition 1.2 The random variable X follows a Gamma distribution with parameters á, b (*Gamma(á,b)*) if its probability density function is given by

$$f(x) = \frac{a^b}{\Gamma(b)} x^{b-1} \exp(-ax), \qquad x > 0, \quad a, b > 0, \tag{1.5}$$

where Ã(á) is the Gamma function defined by

$$\Gamma(a) = \int_0^\infty x^{a-1} \exp(-x) dx \tag{1.6}$$

The parameter á of the Gamma distribution is the scale parameter. It is known that the random variable $Y = cX$ also follows a gamma distribution, with parameters á/c and b.

Definition 1.3 A continuous random variable X follows a Beta Type I distribution with parameters á,b (*ÂetaÉ(á,b)*) if its probability density function is given by

$$f(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, \qquad x > 0, \quad a, b > 0, \tag{1.7}$$

where Â(á,b) is the Beta function defined by

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1} dx \tag{1.8}$$

Definition 1.4 A continuous random variable X follows a Beta type II distribution with parameters á,b (*ÂetaIÉ(á,b)*) if its probability density function is given by

$$f(x) = \frac{x^{a-1}}{B(a,b)(1+x)^{a+b-1}}, \qquad x > 0, \quad a, b > 0, \tag{1.9}$$

It can be verified that the Beta II (á,b) distribution arises as a Gamma mixture of the form

$$Gamma(p,b) \underset{p}{\Lambda} Gamma(1,a)$$

Equivalently, if X_1 and X_2 follow Gamma distributions with parameters (1,a) and (1,b) respectively the random variable $Y = \frac{X_1}{X_2}$ follows a beta II distribution with parameters a and b. This provides an example for Proposition 1.1.

Mixture Models and Compound or Generalised Distributions

Definition 1.5 Consider the random variable S which can be represented as

$$S = X_1 + X_2 + ... + X_N \tag{1.10}$$

where N is a discrete random variable and X_i's are i.i.d. random variables each having a density function f. Let the distribution of N be defined by a probability function p. Then S is said to follow a compound p distribution. The density of the resulting distribution of S is denoted by $p \vee f$. The distribution defined by the density f is called the summand distribution, as it is the distribution of the summands X_i. Some authors use the term generalised p distribution.

Definition 1.6 A compound (or generalised) distribution is called as a compound Poisson distribution if the distribution of N is the Poisson distribution.

An alternative terminology that has been used to refer to these models, is generalised Poisson models. This term has also been used to describe situations other than that in (1.10) and some confusion, as noted by Chatfield and Theobald (1972), may arise. To avoid any confusion in the sequel, we adopt the term compound models to refer to models of the form (1.10).

Another useful connection of mixture models and compound distributions is given in Gurland (1957).

Proposition 1.2 (Gurland, 1957) If a density function g has probability generating function of the form $[\phi(t,a)]^n$ where n is a parameter and $\phi(t,a)$ is a function independent from n, then it holds that

$f \vee g$ is equivalent to $g(x|n) \underset{n}{\Lambda} f(n)$

For example, if we consider a compound Poisson distribution of the form $Poisson \vee Binomial$ we may obtain the same distribution as a binomial mixture.

Proposition 1.3 (Gurland, 1957). It holds that

$$\left[f(x|\lambda) \underset{\lambda}{\wedge} g(\lambda)\right] \vee h \text{ is equivalent to } \left[f \vee h\right] \underset{\lambda}{\wedge} g(\lambda)$$

Definition 1.7 We say that a distribution with probability function p is the convolution of the distributions with probability functions f and g and we denote this by $(f * g)$ if

$$p(x) = \sum_{n=0}^{x} f(x-n)g(n)$$

The convolution is the distribution of the sum $Y = X + Z$, where X follows a distribution with probability function f and Z follows a distribution with probability function g respectively. In the case of continuous random variables X or Z we replace the sum with an integral.

Proposition 1.4 Assuming that the probability density function $g(y|\mu)$ does not depend on λ, it holds that

$\left[f(x|\lambda) * g(y|\mu)\right] \underset{\lambda}{\wedge} h(\lambda)$ is equivalent to $\left[f(x|\lambda) \underset{\lambda}{\wedge} h(\lambda)\right] * g(y|\mu)$

A simple proof of Proposition 1.4 can be obtained by interchanging the order of the integrations involved in the mixture and the convolution.

The above results are simply few of the properties of general mixture models. In the sequel, we will use them for deriving related results. De Vylder (1989) provided some other relations between mixtures and compound distributions.

1.3 The Poisson Distribution

The Poisson distribution can naturally describe phenomena which occur randomly in time or space. It is also known as the law of rare events, since it can be derived from the binomial distribution when the probability of success is small. The range of such events which the Poisson can describe extends over several aspects of our life. Such aspects include

- the number of accidents incurred by an individual in a given time period
- the number of accidents in a given patch of road
- the number of goals scored by a team during a football game
- the number of particles emissioned in an experiment
- the number of plants in a given field
- the number of buys of a product in a given time period, etc.

In fact, such examples are numerous and we just mentioned few of them, trying to include such diverse fields as accident theory, sports, physics, ecology and market research. The common element which allows the Poisson distribution to describe such different situations is the assumed randomness in the occurrence of all the above events.

The term randomness has been misunderstood by many people. In the sequel we will try to formulate the meaning of this word on a mathematical basis.

Definition 1.8 A discrete random variable X is said to follow the Poisson distribution with parameter ë denoted by $Po(\lambda)$ if its probability function is given by

$$P(X = x|\lambda) = P(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!},\ \text{x=0 ,1 , ...,}\ \text{ë>0} \tag{1.11}$$

The Poisson distribution arises from the Poisson process. The Poisson process is a very common stochastic process and standard textbooks for stochastic processes contain sufficient material for it (e.g. Feller,1968). We will briefly describe the Poisson process.

Let N(t) be a random variable denoting the number of events occurred by time t. Also let $P_{mn}(t+dt) = P(N(t+dt) = n \,|\, N(t) = m)$ denote the probability that n events have occurred by time t+dt given that at time t m events had occurred. The interest lies in the transition probabilities $P_{mn}(t+dt)$, because they characterise the process. These probabilities are defined as:

$$P_{mn}(t, t+dt) = \begin{cases} 1 - k_m(t)dt, & n = m \\ k_m(t)dt, & n = m+1 \\ o(dt), & n > m+1 \end{cases} \tag{1.12}$$

where *o(dt)* represents a negligible quantity. The Poisson process is based on the assumption that the infinitesimal risk $k_m(t)$ is constant. This implies that the probability of an event occurring in a very small interval is the same for all intervals, independent of the number of previous events and the position of the interval. For the Poisson process:

$$k_m(t) = \lambda \tag{1.13}$$

Then, from (1.12) and (1.13), the number of events in the entire interval (0,T) follows a Poisson distribution with parameter *ëT*; ë is called the intensity of the process and it reflects how often an event occurs.

It is interesting that the occurrence of an event in an interval does not affect the occurrence of events in any other interval. This implies that, for example, in the case of accidents, the individual will not learn from any event and the probability of incurring a further accident is constant over the entire period. This characterises the notion of randomness. The events occur by pure chance and the times of the occurrence of these events do not play any role in the occurrence of events in subsequent intervals.

The Poisson distribution has been studied in depth by numerous researchers in the last century. A broad review for the Poisson distribution can be found in Johnson *et al.* (1992).

1.4 Related Models

1.4.1 Proneness Model

When we study a population using the Poisson distribution, we assume that all the individuals comprising the entire population have the same intensity parameter. Then we talk about a homogeneous population. How natural is such an assumption in practice? The Poisson distribution assumes homogeneity of the population and this can be a very strict assumption. In practice, it would be more realistic to assume that the population is inhomogeneous. To represent this inhomogeneity we assume that the intensity parameter ë is itself a random variate that varies from individual to individual according to a distribution. Then, by the law of total probability, the distribution of the entire population is a mixed Poisson distribution.

Definition 1.9 A random variable X follows a mixed Poisson distribution with mixing distribution having probability density function g if its probability function is given by

$$P(X = x) = P(x) = \int_0^\infty \frac{e^{-\lambda}\lambda^x}{x!} g(\lambda) d\lambda \quad , \qquad \text{x=0, 1, ...} \tag{1.14}$$

We will denote the mixed Poisson distribution with mixing distribution the distribution with density function g as the *MP(g)* distribution. In the sequel we will use the notation $P(x)$ for the probability function of the mixed Poisson distribution.

The above model is often called the proneness model (see, e.g. Cane, 1975). The parameter ë reflects the proneness of any individual to incur an event. Each individual has its own value for ë and ë has a probability density function $g(\lambda)$. Note that ë is not necessarily a continuous random variable. It can be discrete or it can take a finite number of values. The latter gives rise to finite Poisson mixtures which will be described in the sequel.

The concept of proneness can be extended to all the examples described above, and the term is used in the wider sense. For example, the concept of proneness in market research can be associated with the inherent characteristics of the product which make the product appealing to the consumer. Also proneness in the context of a football game can refer to the inherent characteristics of the team that determine its performance like its composition, technique and style of play etc.

The plausibility of the assumption of inhomogeneity has made the examination of mixture models very useful in practice, since they can describe phenomena for which the simple Poisson model is inadequate. For example, a data set whose variance exceeds the mean can be described better by a proneness model, rather than by a simple Poisson model.

A good review for the proneness model can be found in Arbous and Kerrich (1951, 1954). Chapter 2 is devoted to examining the mixed Poisson distributions defined in (1.11). A variety of mixed Poisson distributions are reviewed and the properties of these distributions are examined in depth.

1.4.2 Contagion Model

The Poisson process was based on the assumption of constant infinitesimal risk in (1.13) over the whole period of observation and of independence between any two events. This assumption is not always realistic. In a variety of cases we may assume that each individual learns from each event in such a way that he (or she) is less probable to incur another event. Arbous and Kerrich (1951) described the least contagious model by assuming that after the first event the infinitesimal risk is reduced and then it remains constant for the remaining period of observation. They called this model as the 'burnt finger' model because it resembles the simple situation where a child touches the fire and learns not to touch it again in his entire life. They derived the observed distribution of events of such a process.

Roughly speaking, the contagion model assumes that each event results in a change on the infinitesimal risk. More formally $k_m(t)$ depends on both m and t. In the sequel, we use the term contagion model to refer to a process in which $k_m(t)$ depends on both m and t, irrespectively of the fact that $k_m(t)$ can be an increasing or a decreasing function of m.

A well known example of such a process is the so-called Polya process for which

$$k_m(t) = \frac{a+m}{b+t} \tag{1.15}$$

The resulting distribution is the negative binomial distribution. McFadden (1965) described a more general process the so called mixed Poisson process (see also Willmot and Sundt,1989 and Grandell, 1997). Every mixed Poisson *MP(g)* distribution can arise from a contagion model if we define as infinitesimal risk the quantity

$$k_m(t) = \frac{\int_0^\infty \lambda^{m+1} e^{-\lambda t} g(\lambda) d\lambda}{\int_0^\infty \lambda^{m} e^{-\lambda t} g(\lambda) d\lambda} \tag{1.16}$$

For t=1, relation (1.16) simplifies to

$$k_m(1) = \frac{(m+1)P(m+1)}{P(m)} \tag{1.17}$$

where *P(m)* is the probability function of the *MP(g)* distribution as given in (1.14).

This has the unfortunate consequence that observing a data set which can be described adequately by a mixed Poisson distribution does not allow us to know from which model they came. At least two models can result in the same mixed Poisson distribution.

Cane (1977) demonstrated the existence of this problem for the negative binomial distribution, the most common mixed Poisson distribution. The negative binomial distribution can also be derived as a compound Poisson distribution. Hence three quite different models can be considered as having led to the same negative binomial namely:

- the proneness model
- the contagion model
- the compound Poisson model

Clearly, with the count data in hand only we are not able to distinguish between these three models. Cane (1977) tried to distinguish between the three models using additional information concerning the exact times of the occurrence of events. Unfortunately, such information did not offer any help in distinguishing between the proneness and the contagion models since both possess the same joint distribution for the time of occurrence of the events. In the same paper Cane (1977) showed that the situation is similar with other mixed Poisson distributions.

A few years later, Xekalaki (1983a) examined the same non-identifiability problem under another mixed Poisson distribution, the generalised Waring distribution. In this case, the joint distribution of the times of the occurrence of accidents did not provide any information for distinguishing between the proneness and the contagion model. Xekalaki (1983a) showed that constructing confidence intervals for the proneness parameters is not helpful either since the intervals are very wide and cannot support any of the assumptions explicitly. However, a third model that assumed spells of accidents for individuals exposed to variable external risk led to a distribution for the times of accidents which differed in form from that of the other two models. Then, the spells model can be identifiable on the basis of such information.

Bates and Neyman (1952a,b) examined the same problem for the first time. They concluded that in the bivariate case the examination of the times of occurrence can provide slight information, which favours one of the two models. Intuitively, we expect that for the proneness model the expected time for the occurrence of a new event remains constant for every individual. On the contrary the expected time for the occurrence of a new event becomes smaller for the contagion model. The problem is how this can be deduced from the data. Wasserman (1983) proposed the use of a likelihood ratio test to test the proneness hypothesis against the contagion hypothesis based on the time of occurrence of the events. No power results are known for this test. Xekalaki (1984a) demonstrated that using the generalised Waring distribution identifiability is possible in the two-dimensional case.

Finding methods for distinguishing between proneness and contagion models still remains an open problem.

1.4.3 Models Based on Both the Contagion and the Proneness Hypotheses

One may consider a synthesis of the two models by constructing a model where proneness and contagion are both present. Such an example is the model constructed by Panaretos (1989). He assumed a contagion model for the diffusion of the surnames, in the sense that the more the surnames in a time period the more there will be in the next time period (contagion). However, every name can have its own parameter of contagion. In other words, every name has a different rate of contagion. The resulting distribution was the Yule distribution which is a special case of the generalised Waring distribution.

1.5 Discussion

In this introductory chapter a brief introduction for mixture models was made. Naturally, mixture models arise as proneness models, i.e. models which take into account the inhomogeneity of the population. Moreover mixture models can also be derived via certain other processes, e.g. contagion process. Even though the problem of determining the process behind the mixture model is an interesting problem connected with the mechanism that generates the data, it will not be pursued any further in this thesis.

The applicability of mixture models in real problems, has led to a vast number of research papers in the statistical literature. The next chapters are devoted to the study of structural properties of mixtures pertaining their use in practice. Chapter 2 contains an extended review on mixed Poisson distributions. According to the choice of the mixing distribution a large number of mixed Poisson distributions can be obtained. Unfortunately, very few of them have been utilised in practice and a small minority of them have been studied in depth. A large number of properties concerning mixed Poisson distributions is reported. Some common elements in their derivation is exploited and new properties are derived.

Chapters 3-5 treat the case of finite Poisson mixtures. Finite Poisson mixtures are useful models with a variety of applications. Their practical value stems from two facts. The first is that they can describe a population consisting of a finite number of subpopulations. The second is that, by assuming a mixed Poisson model, one is able to estimate only a finite mixing distribution, i.e. even if the true mixing distribution is continuous one can only estimate it via a finite step distribution (see, e.g., Laird, 1978).

Chapter 3 concentrates on the estimation of the mixing distribution using the maximum likelihood (hereafter ML) method. An extended review of all the proposed methods is given as well as some interesting new results. For example, the EM algorithm for mixtures is the most promising method for extracting the ML estimates. An improvement of the EM algorithm is proposed based on some intrinsic properties which hold for all the members of the one exponential parameter family. A comparison of several methods for selecting the initial values of the EM algorithm is also made. This chapter focuses also on the ML method applied to mixture models when some additional information is available for some observations. Furthermore, alternative methods for calculating the ML estimates are examined revealing some problems in their application to real datasets.

Moment estimation methods and related variants are treated in Chapter 4. This chapter contains a new approach to the moment method. The efficiency of the method for both large sample sizes based on asymptotic results and small sample sizes based on a simulation experiment is studied. The most important feature of this chapter is the examination of another neglected issue: the existence of the moment estimates. The system of equations for deriving the moment estimates may be intractable. We

report a simulation experiment which reveals that even for large sample sizes the moment estimates do not exist.

A variant of the method with efficient results is also given. This method utilises the zero frequency instead of the third moment. This approach can increase the efficiency when the proportion of zeros is not small.

Chapter 5 contains new material. We propose a robust alternative to the ML method based on the minimum Hellinger distance method of estimation. We derive minimum Hellinger distance estimates of the mixing distribution. These estimates are very efficient for correctly specified models, but at the same time, they are very robust when the assumed model is incorrect. This property makes them competitive alternatives to the standard ML estimates. Inferential methods based on the minimum Hellinger distance method are also derived.

Chapter 6 treats another problem: the one of testing the Poisson assumption against a Poisson mixture alternative. We derive the Hellinger Deviance Test as an alternative to the Likelihood Ratio Test. The Hellinger Deviance Test is powerful and robust at the same time.

Chapter 7 deals with the problem of determining the number of components in a mixture. This is a very interesting and at the same time involved problem. We propose a procedure which uses the Likelihood Ratio Test sequentially, in order to determine the number of components.

Finally, chapter 8 contains a collection of problems which remain open.

In all the chapters we cite results pertaining to their topic not restricting ourselves to Poisson mixtures. Clearly, many of the methods proposed for Poisson mixtures can be extended to cover the case of mixtures of other families, too. All the chapters contain broad reviews so as to help the reader reconstruct the information concerning the work in these areas.

Chapter 2

Mixed Poisson Distributions

2.1 Introduction

The probability function of a mixed Poisson distribution with mixing density g is given in (1.11). Depending on the choice of the mixing distribution *g* a huge number of different mixed Poisson distributions can be constructed.

As already mentioned in the previous chapter, mixed Poisson distributions are adequate to describe overdispersed data sets, for which the simple Poisson model fails. Historically, the derivation of mixed Poisson distributions was originated by Greenwood and Yule (1920) when they considered the negative binomial distribution as a mixture of a Poisson distribution with a Gamma distribution as the mixing distribution. Since then, a large number of mixed Poisson distributions has appeared in the literature. However, very few of them have attracted the interest of applied researchers. The main reason is that often their form is complicated thus discouraging the researcher to use them.

This chapter is devoted to mixed Poisson distributions. In the first part a large number of properties of mixed Poisson distributions is provided, while mixed Poisson distributions are presented in the second part. Certain common elements between these distributions are pointed out.

2.2 Properties of Mixed Poisson Distributions

We start by examining the properties of the mixed Poisson distribution by giving some simple formulas. Let X be a random variable whose distribution is a mixed Poisson distribution. Then the following two results hold:

1. $P(X \le x) = \int_0^\infty \frac{e^{-\lambda}\lambda^x}{x!} G(\lambda) d\lambda$ and

2. $P(X > x) = \int_0^\infty \frac{e^{-\lambda}\lambda^x}{x!} [1 - G(\lambda)] d\lambda$

where $G(\lambda)$ is the distribution function of the random variable ë having probability density function $g(\lambda)$.

Denoting the probability function of the Poisson distribution as $P(X = x|\lambda)$ and with the convention $P(X = -1|\lambda) = 0$ we can see that

$$P(X \le x) = \sum_{n=0}^{x} P(n) = \sum_{n=0}^{x} \int_0^\infty P(X = n|\lambda) g(\lambda) d\lambda = \sum_{n=0}^{x} \int_0^\infty P(X = n|\lambda) G'(\lambda) d\lambda$$

where $P(n)$ has been defined in (1.14). Integrating by parts we obtain

$$P(X \le x) = \sum_{n=0}^{\infty} P(X = n|\lambda) G(\lambda)]_0^\infty - \sum_{n=0}^{x} \int_0^\infty [P(X = n-1|\lambda) - P(X = n|\lambda)] G(\lambda) d\lambda$$

The first term of the right hand side vanishes (see also Holgate, 1970) while the second term interchanging the order of summation and integration leads to

$$P(X \le x) = -\int_0^\infty \sum_{n=0}^{x} [P(X = n-1|\lambda) - P(X = n|\lambda)] G(\lambda) d\lambda = \int_0^\infty P(X = x|\lambda) G(\lambda) d\lambda$$

which completes the proof of 1.

To prove 2 we need only to observe that

$$\int_0^\infty P(X = x|\lambda)[1 - G(\lambda)] d\lambda = \int_0^\infty P(X = x|\lambda) d\lambda - \int_0^\infty P(X = x|\lambda) G(\lambda) d\lambda$$

The first integral equals 1 since

$$\int_0^\infty P(X = x|\lambda) d\lambda = \frac{1}{x!} \int_0^\infty e^{-\lambda}\lambda^x d\lambda = \frac{1}{\Gamma(x+1)} \Gamma(x+1) = 1$$

Hence the right hand side of the last expression reduces to $1 - P(X \le x)$ which proves 2.

Unfortunately, very often the cumulative distribution function of a continuous variable is more complicated than the probability density function, and hence the above results are of limited practical use. However, they are interesting since they relate directly the cumulative

density function of a mixed Poisson distribution with the cumulative density function of the mixing distribution.

2.2.1 Comparison with the Simple Poisson Distribution

Suppose that $P(x)$ is the probability function of a mixed Poisson distribution given in (1.14) and $P(x|m)$ is the probability function of a simple Poisson distribution given in (1.11) with the same mean, say m. Then it holds that:

1) $P(0) \geq P(0|m)$, i.e. the probability of observing a zero value is always higher under the simple Poisson distribution and

2) $\frac{P(1)}{P(0)} \leq \frac{P(1|m)}{P(0|m)} = m$, i.e. than the ratio of the first probability to the zero probability is less than the mean for every mixed Poisson distribution.

Shaked (1980) showed that the function $P(x) - P(x|m)$ has exactly two sign changes of the form + - +, namely that the mixed Poisson distribution gives more probability at 0, and has a longer right tail. This result can be used to test if a mixed Poisson distribution is adequate for describing a dataset. The same holds for other mixtures too.

Shaked (1980) also showed that for every convex function $c(x)$ it holds that

$$\sum c(x)P(x) \geq \sum c(x)P(x|m)$$

(For mixtures of continuous densities summation is replaced by integration).

For example, if $c(x) = (x-m)^2$ the property that the variance of the mixed Poisson is greater than the variance of the simple Poisson is obtained. For $c(x) = (x-m)^4$ we can see that the 4th central moment of the mixed Poisson distribution is greater than the 4th central moment of the simple Poisson distribution. Multivariate extensions of this result are given in Schweder (1982). Another generalisation can be found in Lynch (1988).

2.2.2 The Moments of a Mixed Poisson Distribution

We saw in (1.3) that the moments of any mixed distribution can be obtained by weighting with respect to the mixing distribution. Returning to the mixed Poisson case, we obtain that the probability generating function *Q(t)* of a *MP(g)* distribution is given by

$$Q(t) = E\left[t^x\right] = \int_0^\infty \exp\left[\lambda(t-1)\right] g(\lambda) d\lambda \tag{2.1}$$

We can see that (2.1) is the moment generating function of the mixing distribution evaluated at the point (t-1). Successive differentiation of the probability generating function and evaluation of the resulting expression at t=1 gives us the factorial moments of the mixed Poisson distribution. However, this procedure is equivalent to differentiating the moment generating function of the mixing distribution and evaluating it at t=0 which leads to the moments about the origin of the mixing distribution. Thus we have shown the following Lemma:

Lemma 2.1 The factorial moments of the mixed Poisson distribution are the same as the moments about the origin of the mixing distribution.

Thus we may equate the moments about the origin $E(X^r)$ of the mixed Poisson distribution to those of the mixing distribution. The first five moments about the origin are given with respect to the first moments about the origin $E(\lambda^r)$ of the mixing distribution as:

$$\begin{aligned}
&E(X) = E(\lambda) \\
&E(X^2) = E(\lambda^2) + E(\lambda) \\
&E(X^3) = E(\lambda^3) + 3E(\lambda^2) + E(\lambda) \\
&E(X^4) = E(\lambda^4) + 6E(\lambda^3) + 7E(\lambda^2) + E(\lambda) \\
&E(X^5) = E(\lambda^5) + 10E(\lambda^4) + 25E(\lambda^3) + 15E(\lambda^2) + E(\lambda)
\end{aligned} \tag{2.2}$$

In particular, we have for the variance of the mixed Poisson distribution that

$$Var(X) = E(X^2) - \left[E(X)\right]^2 = E(\lambda^2) + E(\lambda) - \left[E(\lambda)\right]^2 = E(\lambda) + Var(\lambda) \tag{2.3}$$

From the above result it becomes obvious that the variance of a mixed Poisson variate is always greater than the variance of a simple Poisson variate with the same mean. Molenaar and Van Zwet (1966) gave sufficient and necessary conditions for distributions which also have the same property. See also the relevant works of Schweder (1982), Cox (1983) and Gelfand and Dalal (1990).

From the above equation we can see that the variance of the mixed Poisson distribution can be decomposed into two components: the first component can be attributed to randomness and is the variance of a Poisson variate with mean equal to E(ë) and the second component is the variance imposed by the mixing distribution. It is interesting to note that this scheme is similar to the one used in the well known analysis of variance methods (ANOVA) for normal models. In the ANOVA we decompose the total variance into two components. The first is the component that corresponds to the case of equal means and the second represents the variation imposed by the different means of the subpopulations.

The variance to mean ratio for a mixed Poisson distribution is always greater than 1, which is the value which characterises the Poisson distribution. This property has been used to test the null hypothesis that the data come from a Poisson distribution versus the alternative hypothesis that the data come from a mixed Poisson distribution.

The relationship between the moments of the mixed Poisson distribution and the mixing distribution is useful in estimating the mixing distribution from an observed data set. This is discussed in chapter 4.

2.2.3 The Convolution of Two Mixed Poisson Random Variates

The convolution of a mixed Poisson distribution with mixing density *f*, *MP(f)*, with a mixed Poisson distribution with mixing density *g*, *MP(g)*, is a mixed Poisson distribution with mixing density the convolution of the densities g and f. To show this we observe that the probability function of the convolution of *MP(f)* and *MP(g)* is given by

$$P(x)=\sum_{n=0}^{x}\left(\int_0^\infty \frac{e^{-\lambda}\lambda^n}{n!}f(\lambda)d\lambda\right)\left(\int_0^\infty \frac{e^{-\theta}\theta^{x-n}}{(x-n)!}g(\theta)d\theta\right)=$$

$$=\sum_{n=0}^{x}\int_0^\infty\int_0^\infty \frac{e^{-(\lambda+\theta)}\lambda^n\theta^{x-n}}{n!(x-n)!}f(\lambda)g(\theta)d\theta d\lambda \ .$$

By changing the order of summation and integration and by using the binomial expansion $(a+b)^n=\sum_{x=0}^{n}\binom{x}{n}a^n b^{x-n}$ we obtain

$$P(x)=\int_0^\infty\int_0^\infty \frac{e^{-(\lambda+\theta)}(\lambda+\theta)^x}{x!}f(\lambda)g(\theta)d\theta d\lambda$$

On substituting y=ë+è the above formula reduces to

$$P(x) = \int_0^\infty \frac{e^{-y} y^x}{x!} \left(\int_0^y f(y-\theta) g(\theta) d\theta \right) dy = \int_0^\infty \frac{e^{-y} y^x}{x!} t(y) dy$$

where t(y) is the density of a distribution which is the convolution of f and g.

For an alternative proof see Feller (1943).

Definition 2.1 A distribution with probability density function f is said to be reproductive if the sum of two random variables X_1 and X_2 , each having probability density function *f*, follows itself the distribution with probability density function *f*.

Some well known examples of reproductive distributions are the normal distribution, the exponential distribution and the Poisson distribution, among others.

We now give the following Lemma:

Lemma 2.2 The sum of two mixed Poisson variates *MP(f)* has a *MP(f)* distribution if the distribution f is itself reproductive.

To prove Lemma 2.2, we observe that since the convolution of two mixed Poisson distributions is itself a mixed Poisson distribution with mixing density the convolution of the two mixing densities. Since the distribution defined by f is reproductive the convolution of the mixing densities is also defined by f. Therefore, the mixed Poisson distribution is an *MP(f)* distribution.

Another useful result is the following. Consider the convolution of a *MP(f)* and a Poisson distribution. Since the Poisson distribution can be regarded as an *MP(Ä$_{ë}$)* distribution where Ä$_{ë}$ is the density of the degenerate distribution at the point ë, the above convolution is a mixed Poisson distribution. The mixing density is the convolution of the density *f* with the density $Ä_{ë}$. This is a shifted version of f. (In particular, it is the distribution of the random variable $Y = X + \lambda$, where the density of X is *f*). Thus we saw that the convolution of a *MP(f)* with a Poisson distribution is an *MP*(shifted *f*). The Dellaporte distribution (see, e.g. Ruohonen, 1988) is the distribution of the convolution of a Poisson distribution with a negative binomial distribution. Willmot and Sundt (1989) showed that the Dellaporte distribution is a mixed Poisson distribution with a shifted Gamma as mixing distribution.

2.2.4 Identifiability

The term identifiability of mixtures refers to the ability of identifying the mixing distribution of a given mixed distribution. The identifiability of a mixture is important since identifiability ensures that the mixing distribution characterises the mixed distribution. More formally we say that:

Definition 2.2 Mixtures of the probability function $f(x|\theta)$ are identifiable if and only if $\int f(x|\theta)h_1(\theta)d\theta = \int f(x|\theta)h_2(\theta)d\theta$ implies that $h_1(\theta) = h_2(\theta)$ for all the values of è. In the case of discrete mixtures integration is replaced by summation.

In our case, $f(x|\theta)$ is the probability function of a Poisson distribution. Mixtures of the Poisson distribution (finite or not) are identifiable. This means that every mixed distribution corresponds to one and only one mixing distribution. The identifiability of Poisson mixtures (finite or not) was first proved by Feller (1943) who pointed out that the probability generating function of a mixed Poisson distribution is the Laplace transform of the mixing distribution. It is known that the Laplace transform of any function is unique. Hence the probability generating function of any mixed Poisson distribution is unique. Since the probability generating function uniquely determines the distribution it follows that Poisson mixtures are identifiable. Later, Teicher (1961) showed that mixtures on n of distributions with probability generating function of the form $(h(t))^n$ are identifiable. The Poisson distribution belongs to this family, which also contains the normal, the gamma and the binomial distributions among others. The identifiability of Poisson mixtures has also been examined in Xekalaki and Panaretos (1983), Xekalaki (1985) and Lindsay and Roeder (1993). The property of identifiability is interesting since only in this case it is sensible to estimate the mixing distribution. Related material can be found in Barndorff-Nielsen (1965), Tallis (1969) and Yakowitz and Spragins (1969). Identifiability of finite mixtures has also been discussed by Teicher (1963) and Al-Hussaini and El-Dab (1981).

Chen (1995) established the notion of strong identifiability, which is satisfied by the Poisson mixtures. He uses this notion to prove the rate of convergence of any estimator of the mixing distribution to the true mixing distribution.

2.2.5 Modality

Holgate (1970) showed that a *MP(g)* distribution is unimodal if g is unimodal. Note that he used the term unimodal to refer to distributions with one mode or with several modes at successive points. So the unimodality of a mixed Poisson distribution depends on the unimodality of its mixing distribution. This result holds only if g is absolutely continuous. A counterexample for the case of discrete mixing distributions is the Neyman distribution which is a mixed Poisson distribution with a Poisson distribution as the mixing distribution. The Neyman distribution is known to be multimodal (see Douglas, 1980) even though the Poisson distribution is unimodal.

Bertin and Theodoreskou (1995) extended the results of Holgate to the case of not absolutely continuous mixing distributions. In general, modality for mixture models is considered in Kemperman (1991).

2.2.6 Infinite Divisibility

A random variable X is said to be infinitely divisible if its characteristic function ö(t) is such that $\left[\phi(t)\right]^{1/n}$ is itself a characteristic function. In other words, a distribution is infinitely divisible if it can be written as the distribution of the sum of n independently and identically distributed random variables. The simple Poisson distribution is an example, since the sum of n independent Poisson variates is itself a Poisson variate (with different parameter of course). A result concerning the mixed Poisson distribution was given by Douglas (1980):

Proposition 2.1 (Douglas, 1980). If the mixing distribution is infinitely divisible then the mixed Poisson distribution is infinitely divisible, too.

Feller (1968) showed another interesting result concerning infinitely divisible discrete distributions.

Lemma 2.3 (Feller, 1968). A discrete infinitely divisible distribution can be obtained also as a compound Poisson distribution.

Hence, a mixed Poisson distribution which is infinitely divisible can also be represented as a compound Poisson distribution. This implies that its probability generating function can be written either as

$$\int_0^\infty e^{\lambda(s-1)} f(\lambda) d\lambda$$

with $f(\lambda)$ the probability density function of the mixing distribution or as

$$e^{\lambda(Q(s)-1)}$$

with $Q(s)$ the probability generating function of a well defined distribution (the summand distribution).

Well known examples are the negative binomial distribution (which will be described in the sequel), the Poisson- inverse Gaussian distribution, and the generalised Waring distribution. Note that for the two first cases the form of the summand distribution is known while for the latter case the form of the summand distribution has not been derived in a closed form.

2.2.7 Mixed Poisson and Compound Poisson Distributions

The infinite divisibility of some mixed Poisson distributions implies their representation as compound Poisson distributions. A proof of this result can be found in Ospina and Gerber (1987).

As noted above, a compound Poisson distribution has probability generating function $G(z)$ of the form

$$G(z) = \exp\left[\lambda\left(Q(z) - 1\right)\right] \tag{2.4}$$

where $Q(z)$ is the probability generating function of the summand distribution. Our aim is to identify the summand distribution which allows the compound Poisson representation of a mixed Poisson distribution. If we solve the above equation we find that the probability generating function of the summand is

$$Q(z) = \frac{\ln G(z)}{\lambda} + 1 \tag{2.5}$$

Therefore the probability function of the summand can be obtained by successive differentiation of (2.5). In practice, this does not always lead to a closed form for the probability function of the summand distribution. Of course, in all cases we are able to calculate numerically the probability function, using the following result due to Panjer (1981)

who showed that compound Poisson distributions can be calculated via recursive schemes. In the case where the summand distribution is discrete with probability function, say $f(x)$, the probability function of corresponding compound Poisson distribution, say $g(x)$, can be obtained recursively via the following:

$$g(x) = \sum_{y=1}^{x} \frac{\lambda y}{x} f(y) g(x-y) \tag{2.6}$$

with

$$g(0) = \exp\left[-\lambda + \lambda f(0)\right] \quad . \tag{2.7}$$

From (2.6) we are able to derive the probability function of a compound Poisson distribution from the form of the summand distribution. For the converse result relations (2.6) and (2.7) have to be solved for $f(x)$. We thus obtain,

$$f(0) = \frac{\ln g(0) + \lambda}{\lambda} \quad , \qquad f(1) = \frac{g(1)}{\lambda g(0)}$$

and $$f(x) = \frac{g(x)}{\lambda g(0)} - \frac{1}{x g(0)} \sum_{y=1}^{x-1} y f(y) g(x-y) \qquad \text{for x=2, 3, ...} \tag{2.8}$$

Willmot (1986) proposed choosing the value of ë by imposing the condition Q(0)=0. However, we can verify that the successive ratios of the form $\frac{f(x+1)}{f(x)}$ do not depend on ë, apart form the ratio $\frac{f(1)}{f(0)}$. It should be noted that the above scheme applies only to the case of discrete summand distributions and that it cannot be used for estimation purposes. If we try to use the empirical relative frequencies as an estimate of $g(x)$ then we may obtain unacceptable values for the probability function of the summand, i.e. negative values or values greater than 1. Then, the whole scheme fails.

2.2.8 Posterior Moments of ë

We will give the following Proposition concerning the posterior expectation of the random variable ë:

Proposition 2.2 Suppose that X follows a *MP(g)* distribution. Then the posterior expectation $E(\lambda^r | X = x)$ is given by:

$$E\left(\lambda^r \mid X=x\right)=\frac{P(x+r)}{P(x)}(x+1)\ldots(x+r)$$

where P(x) is the probability function of a *MP(g)* distribution.

Proof.

The posterior expectation $E\left(\lambda^r \mid X=x\right)$ is expressed as

$$E\left(\lambda^r \mid X=x\right)=\frac{\frac{1}{x!}\int_0^\infty e^{-\lambda}\lambda^{x+r}g(\lambda)d\lambda}{\frac{1}{x!}\int_0^\infty e^{-\lambda}\lambda^{x}g(\lambda)d\lambda}=\frac{\frac{(x+r)!}{x!}\int_0^\infty \frac{e^{-\lambda}\lambda^{x+r}}{(x+r)!}g(\lambda)d\lambda}{P(x)}=\frac{P(x+r)}{P(x)}(x+1)\ldots(x+r)$$

Note that the above results may be extended to the case of negative r whenever $(x+r)>0$. This enables one to find, for example, posterior expectations of the form $E\left(\frac{1}{\lambda}|X=x\right)$.

Johnson (1957) showed that the posterior first moment of ë is linear if and only if the mixing distribution is the Gamma distribution. Johnson (1967) generalised this result to show that the form of the posterior moment of ë determines the mixing distribution. Nichols and Tsokos (1972) derived more general formulas for a variety of distributions. The results of Cressie (1982) are also pertinent. More recently Sapatinas (1995) gave the special forms of this posterior expectation for several mixed Poisson distributions as well as for other power series mixtures.

It is interesting to note that since the posterior expectation is expressed through the ratio $P(x+1)/P(x)$ characterises the mixed Poisson distribution among all the mixed Poisson distributions (see also Papageorgiou and Wesolovski, 1997) . Ord (1967) showed that for some basic discrete distributions the ratio $\frac{(x+1)P(X=x+1)}{P(X=x)}$ can provide useful information concerning the distributional form of the population from which the data come. Unfortunately, the practicality of this result is limited since many mixed Poisson distributions can have very similar graphs for the quantity $\left(x,\frac{(x+1)P(X=x+1)}{P(X=x)}\right)$ making the identification very difficult.

Bhattacharya (1967) showed the following result in the context of accident theory: if the mixing distribution is a Gamma distribution then the selection in the second time period of individuals with no accidents in the first period will reduce the expected number of accidents in the second period. In particular, he showed that if X and Y are the numbers of accidents in the first and the second period respectively, then $\frac{E(Y)}{E(Y|X=0)} \geq 1$. He also showed that this result is valid for the Poisson-confluent hypergeometric series distribution (Bhattacharya, 1966).

Haight (1965) derived the distribution of the number of accidents in a second period given the removal of persons with more than n accidents in the first period for the case of a negative binomial accident distribution.

2.2.9 Numerical Approximation for the Probability Function of a Mixed Poisson Distribution

For many mixed Poisson distributions the direct calculation of probabilities is involved and some numerical methods need to be used. Even the use of powerful recursive schemes require direct calculation of initial values. In the sequel, we present some numerical methods for the efficient calculation of the probabilities.

a) Taylor expansions

The first method can be found in Ong (1995) and it is based on a Taylor expansion of a special function of a gamma variate:

Lemma 2.4 (Ong, 1995). Let $g(\lambda)$ be the probability density function of the mixing distribution of a mixed Poisson distribution. If $g(\lambda)$ has a finite n-th derivative at the point k, the probability function *P(k)* of the mixed Poisson distribution has the formal expansion:

$$P(X=k) = g(k) + \frac{1}{k}\sum_{y=2}^{n} \frac{\mu_y h^{(y)}(k)}{y!},$$

where $h(k) = kg(k)$, $h^{(i)}(k)$ denotes the i-th derivative of h(k) with respect to k and μ_y is the y-th moment about the mean of a gamma random variable with scale parameter equal to 1 and shape parameters equal to k.

The above approximation has some disadvantages. The first is that we cannot obtain *P(0)*. On the other hand, we need to evaluate the derivatives of the mixing distribution, which, even if they exist, is a very tedious task.

We now give another easier approximation based on the Taylor expansion of the probability function of a mixed Poisson distribution.

From the probability function of a *MP(g)* we have that

$$P(X = x) = \frac{1}{x!}\int_0^\infty \exp(-\lambda)\lambda^x g(\lambda)d\lambda = \frac{1}{x!}E_g\left[e^{-\lambda}\lambda^x\right] \tag{2.9}$$

Relationship (2.9) shows that the probability function of a mixed Poisson distribution is the expectation of a certain function of ë with respect to the mixing distribution. Using the standard Taylor expansion for a function of a random variable we can obtain the probability function of the mixed Poisson distribution. This method requires only the moments of the mixing distribution and not its probability density function. Hence, the probability function of the mixed Poisson distribution can be approximated by:

$$P(X = x) = \frac{1}{x!}\left(t(\mu) + \sum_{r=1}^{\infty}\frac{\mu_r t^{(r)}(\mu)}{r!}\right), \tag{2.10}$$

where $ì_r$ is the r-th central moment of the mixing distribution, $ì=ì_1$ i.e. the mean, $t(\mu) = \exp(-\mu)\mu^x$ and $t^{(r)}(\mu)$ is the r-th derivative of t evaluated at ì. The derivatives of t(ì) can be easily obtained from the above iterative scheme, thus:

$$\begin{aligned} t^{(0)}(\theta) &= \exp(-\theta)\theta^x, \\ t^{(n+1)}(\theta) &= \sum_{i=0}^{n}\binom{n}{i}t^{(i)}(\theta)h^{(n-i)}(\theta) \quad . \end{aligned} \tag{2.11}$$

Here, the function h(è) has derivatives given by

$$h^{(0)}(\theta) = \left(\frac{x}{\theta} - 1\right), h^{(1)}(\theta) = -\frac{x}{\theta^2}, h^{(n+1)}(\theta) = -\frac{(n+1)h^{(n)}(\theta)}{\theta} \quad . \tag{2.12}$$

Hence, using (2.11) and (2.12) we can calculate (2.10). This Taylor expansion provides very ‘accurate’ values for the probabilities. The limitation of the above formula is that it requires the existence of the moments about the mean of the mixing distribution.

However, for some mixed Poisson distributions for which the calculation of the probability function is tedious but their moments are known to exist (e.g., lognormal) this is useful.

b) The Probability Function of the Mixed Poisson Distribution as an Infinite Series Involving the Moments of the Mixing Distribution

Another useful formula which relates the probability function of a mixed Poisson distribution with the moments of the mixing distribution can be used. In particular, we have that the probability function can be written as an infinite series expansion using the moments about the origin of the mixing distribution.

Lemma 2.5 The probability function of a mixed Poisson distribution can be written in the following form given that the moments of è exist:

$$P(X=x)=\frac{1}{x!}\sum_{r=0}^{\infty}\frac{(-1)^r}{r!}\mu_{x+r}(\theta)$$

where $\grave{\imath}_r$ (è) is the r-th moment about the origin of è.

Proof

We have for the probability function of the mixed Poisson distribution that

$$P(X=x)=\frac{1}{x!}\int_0^{\infty}\exp(-\theta)\ \theta^x g(\theta)d\theta\ =\ \int_0^{\infty}\left(\sum_{r=0}^{\infty}\frac{(-\theta)^r}{r!}\right)\frac{\theta^x}{x!}g(\theta)\ d\theta$$

Interchanging the order of summation and integration we obtain

$$P(X=x)=\sum_{r=0}^{\infty}\int_0^{\infty}\frac{\theta^{x+r}(-1)^r}{x!\,r!}g(\theta)d\theta=\sum_{r=0}^{\infty}\frac{(-1)^r}{x!\,r!}\int_0^{\infty}\theta^{x+r}g(\theta)d\theta\,.$$

Hence the result.

The above result can alternatively be seen as follows:

It is obvious by their definition that the probability generating function *G(t)* of a mixed Poisson distribution and the moment generating function *M(t)* of the mixing distribution (if it exists) are related, thus

$$G(t)=\int_0^{\infty}e^{\lambda(t-1)}f(\lambda)d\lambda\ \ =\ E(e^{\lambda(t-1)})\ =\ M(t-1)$$

It is also known that the probability function of a discrete distribution with probability generating function *G(t)* is written as

$$P(X=x)=\frac{1}{x!}G^{(x)}(0), \tag{2.13}$$

where $G^{(x)}(t)$ is the x-th derivative of G(t) evaluated at the point t. Furthermore, for a mixed Poisson distribution the following relation holds:

$$G^{(x)}(t)=M^{(x)}(t-1), \tag{2.14}$$

while by definition the moment generating function of a random variable can be written as

$$M(t)=\sum_{r=0}^{\infty}\frac{\mu_r}{r!}t^r \tag{2.15}$$

From (2.15) we have that

$$M^{(n)}(t)=\sum_{r=n}^{\infty}\frac{\mu_r}{(r-n)!}t^{r-n} \Rightarrow M^{(n)}(-1)=\sum_{r=n}^{\infty}\frac{\mu_r}{(r-n)!}(-1)^{r-n} \Leftrightarrow .$$

$$\Leftrightarrow M^{(n)}(-1)=\sum_{r=0}^{\infty}\frac{\mu_{r+n}}{r!}(-1)^r \tag{2.16}$$

Using (2.14) and (2.16), equation (2.13) can be written as

$P(X=x)=\frac{1}{x!}M^{(x)}(-1)=\frac{1}{x!}\sum_{r=0}^{\infty}\frac{\mu_{r+x}}{r!}(-1)^r$ which is the desired result.

Katti (1966) derived a similar result for compound distributions.

c) Gauss-Laguerre Polynomials

This approximation is based on the adjusted Laguerre polynomials or simply Gauss-Laguerre polynomials. Following Press et al (1992) certain integrals can be approximated using certain weight functions of the integrand evaluated at certain points, thus

$$\int_0^{\infty}e^{-x}x^a f(x)dx=\sum_{j=1}^{n}w_j f(x_j) \tag{2.17}$$

In this formula w_j and x_j , $j=1, \ldots, n,$ are the Gauss-Laguerre weights and abscissas calculated using the methods described in Press *et al.* (1992) and n is the number of points used for the approximation. The authors also gave routines for calculating them. Clearly, the probability function of every mixed Poisson distribution can be approximated using the above formula. It is interesting that using (2.17) the probability function of a mixed Poisson distribution is calculated as a finite mixture of the mixing distribution.

d) Recursive Relations for Mixed Poisson Distributions

Approximating all the probabilities of a probability function is not a good strategy, mainly because of the computational effort required. Willmot (1993) showed that, for several mixed Poisson distributions a recursive formula can be obtained. More specifically, for a mixing density $g(\theta)$ which satisfies the relationship

$$\frac{d\ln g(\theta)}{d\theta} = \frac{\sum_{i=0}^{k} s_i \theta^i}{\sum_{i=0}^{k} w_i \theta^i} ,$$

a recursive formula can be found. If the support of è is (0,+∞) this recursive formula is

$$\sum_{n=-1}^{k} \{\varphi_n + m w_{n+1}\} (m+n)^{(n)} P(m+n) = 0 \qquad , \qquad (2.18)$$

where $a^{(b)} = a(a+1)\ldots(a+b+1)$ and $\phi_n = s_n + (n+1)w_{n+1} + w_n$ with $\phi_{-1} = 0$. Appropriate modifications are available in Willmot (1993) for different supports of è. For this iterative scheme we need only to calculate the first k probabilities. Ong (1995) proposed that the above iterative scheme can be used without exact evaluation of any probability. The idea is to start from a point n at the tail of the distribution putting arbitrarily $P(n) = 1$ and $P(n+1) = 0$. Then by using the above recurrence we may obtain the values $P(n-1)$ up to $P(0)$. Rescaling so that the obtained series sums to 1 leads to the probability function. It is useful to start with a value of n, such that the true $P(n)$ is negligible. We have to note that the recursion described is unstable and should therefore be used with caution.

The above defined recursive scheme led to the increase of the applicability of several mixed Poisson distributions. Earlier, the difficulties in evaluating the probabilities prevented the researchers to use many of the mixed Poisson distributions. Wang and Panjer (1993) criticised the relations with respect to their stability. According to Wang and Panjer (1993) the recurrence relations might be quite unstable, mainly because of the negative coefficients of some probabilities. Then, they proposed to use as starting points for the relations those points where the instability occurs. These points are the points with a negative coefficient in the recurrence representation. They also included several examples.

2.2.10 Simulation Based Reconstruction

An approximate way to construct the probability function for all the mixed Poisson distributions is via simulation. The simulation from a mixed Poisson distribution is easy via the following scheme:

Step 1. Generate *ë* from the mixing distribution.

Step 2. Generate *X* from the *Po(ë)* distribution.

Hence, if we simulate a very large number of values an approximation of the probability function can be obtained. Note that the speed of the simulation depends on the speed of generating a random variate from the mixing distribution. As the number of replications increases, the approximate probability function tends to the true probability function. We can simulate from a mixed Poisson distribution even when we do not know its probability function.

2.2.11 Weighting a Mixed Poisson Distribution

When an investigator records an observation, by nature, according to a certain stochastic model, the recorded observation may not have the original distribution unless every observation is given an equal chance of being recorded. In all other cases where there is a probability w(x) of observing an observation with the value x, the observed distribution differs from the assumed. Patil and Rao (1978) used the term weighted distributions for the distributions arising under such a model, when the weight function $w(x)$ is the probability of observing the value x. Such weighted distributions stem from several different schemes and choices of the weight function, which may not represent a probability. For example when $w(x)$ is proportional to the value x, we assume that the higher the count the more probable to observe it. The value $x = 0$ is never observed. Patil and Rao (1978) used the term visibility bias to describe such models in the context of ecological applications. These models are also called size biased models, and the resulting distributions are called size biased distributions. In general, if the weighted function is $w(x)$ then the probability density function of the weighted distribution obtained from the original density $f(x)$ is given by

$$f_w(x) = \frac{w(x)f(x)}{E[w(x)]}$$

The size biased distribution corresponds to the case where $w(x) = x$, and its probability density function is given by

$$f_w(x) = \frac{xf(x)}{E(x)} \tag{2.19}$$

Patil and Rao (1978) also discussed the effect of *w(x)* in several widely used distributions, while Patil *et al.* (1986), gave some result for discrete distributions as well as the form of the resulting weighted distribution for some discrete distributions. It is interesting to mention the following result concerning mixed Poisson distributions:

Lemma 2.6 A size biased *MP(g)* distribution arises also as a mixture of a size biased Poisson distribution with mixing distribution the size biased version of the original mixing density g.

Proof

The size biased version of the mixing distribution, say $g^*(\lambda)$ has probability density function

$$g^*(\lambda) = \frac{\lambda g(\lambda)}{E(\lambda)} \tag{2.20}$$

while the size biased Poisson distribution has probability function of the form

$$f^*(x|\lambda) = \frac{x}{\lambda}\frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!} \tag{2.21}$$

and then, using (2.20) and (2.21), the mixed size biased Poisson distribution with mixing distribution g^* has probability function

$$f(x) = \int_0^\infty f^*(x)g^*(\lambda)d\lambda = \int_0^\infty \frac{x}{\lambda}\frac{e^{-\lambda}\lambda^x}{x!}\frac{\lambda}{E(\lambda)}g(\lambda)d\lambda =$$

$$= \frac{xP(x)}{E(\lambda)} = \frac{xP(x)}{E(x)}$$

which is (from (2.19)) the size biased version of a *MP(g)* distribution.

For example the size biased negative binomial distribution (the mixing distribution is the Gamma distribution) can be obtained as a mixed size biased Poisson distribution with the size biased Gamma distribution as the mixing distribution.

Seshadri (1991) has shown that mixtures of the length biased Poisson distribution are in fact mixtures with components the simple Poisson distribution and the length biased Poisson distribution.

2.2.12 Shape

The probability function of a *MP(g)* distribution resembles the probability density function of the mixing distribution with adequate adjusted parameters. This fact has been used for approximating the probability function of some mixed Poisson distributions. For example, Best and Gipps (1974) proposed the use of the cumulative distribution function of a Gamma distribution as an approximation to the cumulative distribution function of the negative binomial distribution. The resemblance is much greater for larger values of the mean. If the mean is small, there is a high probability of observing a 0 value, i.e. $P(0)$ is large. This is not true for many continuous densities, and thus the approximation is poor. Cassie (1964) discussed the use of the lognormal distribution instead of the Poisson-Lognormal distribution. Willmot (1989) examined the asymptotic tail behaviour of some mixed Poisson distributions. He showed that the tails of some mixed Poisson distributions look like the tails of their mixing distributions, and he proposed the approximation at the tails by the more tractable continuous mixing distributions. A similar result is given in Perline (1998).

We can also verify that

Lemma 2.7 Grandell (1997). For two mixed Poisson distributions, say $MP(g_1)$ and $MP(g_2)$, we have that $MP(g_1) \rightarrow MP(g_2)$ if and only if $g_1 \rightarrow g_2$ where $\rightarrow$ denotes convergence in distribution.

To verify this result we only need to recall the fact that the probability generating function of a *MP(g)* is the moment generating function of the mixing distribution g. Since the probability generating function and the moment generating function of any distribution characterises the distribution the result follows.

For example, we know that the Beta distribution tends to the Gamma distribution under certain conditions for the parameters are satisfied. Hence, the negative binomial distribution is a limiting case of the Poisson-Beta distribution.

Chen (1995) established the best possible rate of convergence for estimating the mixing distribution in finite mixture models, including the Poisson mixtures. This rate is only $n^{-1/4}$ when the number of support points is not known a priori and only some minimum distance estimators can achieve it.

Adell and de la Cal (1993) , obtained, under fairly general assumptions, the exact order of convergence, in the L_1 distance, of a mixed Poisson distribution to its mixing distribution. Hall (1979) and Pfeifer (1987) discussed the distance between the mixing and the mixed distributions. Lynch (1988) showed that the form of the mixing distribution carries over to the mixed distribution.

2.2.13 Compound Mixed Poisson Distributions

In actuarial applications the distribution of the number of claims can be modelled by a mixed Poisson distribution because it is very common that the population under investigation is not homogeneous. Depending on the distribution of the claim size, the total amount paid by the insurance company follows a compound mixed Poisson distribution. The probability density function of the total amount is usually hard to derive and to compute. So recursive relations are very useful. The most known mixed Poisson distribution, the negative binomial distribution, has been treated in the fundamental paper of Panjer (1981) as it is the only member of the family of mixed Poison distributions with linear first order recurrence relations. Compound mixed Poisson distributions are discussed in detail in a series of papers, by Hesselager (1994a,1996) and Wang and Sobrero (1994). The idea is to construct recurrence schemes based on the recurrence relation for the probabilities of the mixed Poisson distribution. In these papers several examples are included for many members of the family of mixed Poisson distributions.

2.2.14 Mixed Poisson Distributions Arising from the Mixed Poisson Process

As seen before, all the mixed Poisson distributions can be obtained via a mixed Poisson process defined from the infinitesimal risk given in (1.16). A mixed Poisson distribution arising from the mixed Poisson process with infinitesimal risk as in (1.16) is of the form

$$Poisson(t\lambda) \underset{\lambda}{\wedge} g(\lambda) \qquad \text{(Model 1)}$$

where t is the period of observation. In general such models lead to mixed Poisson distributions that differ from those obtained from the model

$$Poisson(\lambda) \underset{\lambda}{\wedge} g(\lambda) \qquad \text{(Model 2)}$$

The resulting distributions coincide for t=1. Our main interest lies on the second model. However, in some circumstances it is of interest to consider the first model, especially when t represents some time period. If we assume that the observed time period is the unit time period, then t=1 and the two models are the same. On the other hand, it is often interesting to consider the Model 1 and how we can relate directly this more general model to Model 2 which is the commonly used model in practice. Note also that several mixed Poisson regression models use Model 1. In these models the Poisson parameter λ is treated as a regressor depending on a series of covariates, while t is a random variable having its own probability function, which is called the overdispersion parameter. For more details on mixed Poisson regression models, one can refer to Lawless (1987), Dean *et al.* (1989), Xue and Deddens (1992), McNeney and Petkau (1994), Wang *et al.* (1996) and Chen and Ahn (1996). Moreover it is worth mentioning that mixed Poisson regression models allow for different variance to the mean relationships offering a wide range of different model for real applications (see, e.g. , Hinde and Demetrio, 1998).

Using definition 2.1 for reproductive distributions we can see that if the mixing distribution is reproductive a rescaling of the random variable does not affect the distributional form of the resulting mixed Poisson distribution but does affect the parameters. In this case the probabilities of the mixed Poisson distribution are easily obtainable. The gamma and the inverse Gaussian distributions are some well known examples of reproductive distributions commonly used as mixing distributions. A problem arises when the mixing distribution is not reproductive.

We will distinguish between the cases $t<1$ and $t>1$.

a) The case when t < 1

When t<1, the Poisson distribution with parameter $t\lambda$ can be represented as a compound Poisson distribution with a Bernoulli distribution as the summand distribution. It is known (e.g. Douglas, 1980) that for compound distributions the probability generating function of the compound distribution is the probability generating function of the simple distribution evaluated at the probability generating function of the summand distribution. More formally if *G(z), S(z)* and *Q(z)* are the probability generating functions of the compound, the simple and the summand distributions respectively we have that

$$G(z) = S\big(Q(z)\big)$$

which for the compound Poisson case reduces to (2.4)

Assume that $Q(z) = q + pz$, *q=1-p*, i.e. the probability generating function of the Bernoulli distribution, where *p* represents the probability of success. Then, the resulting compound Poisson distribution has probability generating function given by

$$G(z) = \exp\big[\lambda p\big(z-1\big)\big]$$

i.e. a Poisson distribution with parameter pë. According to the notation introduced above we may write the model as:

$$Poisson(\lambda) \vee Bernoulli(p)$$

Using this result we can rewrite the model $Poisson(p\lambda) \underset{\lambda}{\Lambda} g(\lambda)$ as

$$\big[Poisson(\lambda) \vee Bernoulli(p)\big] \underset{\lambda}{\Lambda} g(\lambda)$$

and hence (from Proposition 1.3) as

$$\Big[Poisson(\lambda) \underset{\lambda}{\Lambda} g(\lambda)\Big] \vee Bernoulli(p),$$

This shows that the model can be represented as a compound mixed Poisson distribution. Standard techniques for recursive evaluation of the probabilities for compound mixed Poisson distributions are known (see, e.g., Hesselager, 1994a,1996). So the calculation of the probabilities is easy (though tedious in some circumstances).

b) The case when t > 1

If t > 1 the compound Poisson representation of the simple Poisson distribution fails and no results are available for facilitating the evaluation of the probabilities. For this very interesting case there are not known results which can help in calculating the probabilities.

Deriving helpful relations for calculating the probabilities in this case remains an open problem.

2.3 Mixed Poisson Distributions

The aim of this section is to present and bring together a large number of mixed Poisson distributions which have appeared in the literature. We will consider only distributions with a continuous mixing distribution. Discrete mixing distributions have also been considered in the literature, but they will not be treated in detail. As will be seen a large number of mixed Poisson distributions exist. However, their use is limited and only a few of them have been studied in depth and have well understood properties. Further research for classifying and comparing all these distributions is necessary. We will restrict ourselves to presenting the distributions. Numerical techniques are in some cases necessary for the calculation of the probability function, combined with a recursive scheme. The moment estimates of all these distributions are usually straightforward, because of the property of the mixed Poisson distribution which relates the moments of the mixing distribution with those of the mixed distribution.

2.3.1 The Negative Binomial Distribution

The negative binomial distribution is the best known mixed Poisson distribution, and a widely used discrete distribution. Its original derivation as a mixed Poisson distribution was obtained by Greenwood and Yule (1920) by letting the parameter *è* of a Poisson distribution to follow a Gamma distribution.

If we assume that è follows a Gamma distribution with parameters á and b, given in (1.5), then X follows a negative binomial distribution with parameters á and b with probability function given by

$$P(x)=\frac{\Gamma(x+b)}{x!\Gamma(b)}\left(\frac{1}{1+a}\right)^{x}\left(\frac{a}{1+a}\right)^{b} \qquad \text{x=0, 1, ..., á , b >0} \quad . \qquad (2.22)$$

A simple recursive relation for calculating the probabilities is:

$$P(x+1)=\frac{x+b}{x+1}\left(\frac{1}{1+a}\right)P(x) \quad \text{for x=0,1,....}$$

starting with $P(0)=\left(\frac{a}{1+a}\right)^{b}$.

The negative binomial is the only mixed Poisson distribution for which the ratio $\frac{(x+1)P(x+1)}{P(x)}$ is linear with respect to x. This may be used as a quick check of whether the assumption of a negative binomial distribution is plausible.

We note that (2.22) is only one out of several different parameterisations of the negative binomial distribution which flourish in the literature. It can be easily shown that $E(X)=\frac{b}{a}$ and $Var(X)=\frac{b}{a}+\frac{b}{a^2}$. The negative binomial distribution has been used as an alternative to the Poisson distribution when some overdispersion is present in the dataset. The tractability of its probability function before the massive use of computer was the main reason for the use of this distribution to describe overdispersed data in many applications and in a variety of scientific fields. Moment estimates are easily derived as in all the mixed Poisson distributions. However they are not efficient (Sichel, 1951) or they may not exist. ML estimation has been described in Ross and Preece (1985) and Piegorsch (1990) among others.

As will be seen in the sequel, the negative binomial distribution also arises as a special case of some other mixed Poisson distributions. Its fundamental role in the theory of Poisson mixtures led many authors to introduce various generalisations of it.

A lot of research has been carried out for the negative binomial distribution. In Johnson *et al.* (1992) the reader can find a long review for the negative binomial distribution.

If b=1, then the mixing distribution is a *Gamma (á,1)* distribution which is the exponential distribution with probability density function

$$g(\theta)=ae^{-\theta a} \quad , \text{á,è}>0 \tag{2.23}$$

Then, the resulting mixed Poisson distribution is the geometric distribution with probability function given by

$$P(x)=\left(\frac{a}{1+a}\right)\left(\frac{1}{1+a}\right)^{x}, \qquad \text{x=0, 1, . . . , á>0.} \tag{2.24}$$

The geometric distribution is also a well examined discrete distribution. It is always J-shaped with mode at 0.

The negative binomial is an infinitely divisible distribution and hence it can be written as a compound Poisson distribution. Quenouille (1949) showed that if the summand distribution is the logarithmic distribution with probability function given by

$$P(x)=\begin{cases}\dfrac{(1-a)\theta^{x}}{-x\ln(1-\theta)} & ,x=1,2,\ldots\\[2ex] a & ,x=0\end{cases}$$

where $0\le\theta,a<1$, the resulting compound Poisson distribution is the negative binomial distribution. Note that the logarithmic distribution described in Johnson *et al.* (1992) has support to the positive integers, i.e. á=0. The negative binomial can also be derived via the so-called Polya process. This is a mixed Poisson process with infinitesimal risk given in (1.15).

2.3.2 The Poisson-Lindlay Distribution

Sankaran (1970) proposed the Poisson-Lindlay distribution for the analysis of count data. The distribution arises from the simple Poisson distribution if the parameter è follows the Lindlay distribution having probability density function

$$g(\theta)=\frac{p^{2}}{p+1}(\theta+1)e^{-\theta p} \qquad , \qquad \text{è,p>0.}$$

The resulting Poisson-Lindlay distribution has probability function given by

$$P(x)=\frac{p^{2}(p+2+x)}{(p+1)^{x+3}}, \qquad \text{x=0, 1, ..., p>0}$$

The mean and the variance of the Poisson-Lindlay distribution are given by

$$E(X)=\frac{(p+2)}{p(p+1)} \text{ and } Var(X)=\frac{p^{3}+4p^{2}+6p+2}{p^{2}(p+1)^{2}} .$$

A simple recursive formula for calculating the probabilities is the following

$$P(x+1)=P(x)\frac{(p+3+x)}{(p+1)(p+2+x)} \text{ with } P(0)=\frac{p^{2}(p+2)}{(p+1)^{3}} .$$

Despite its simplicity, this distribution has not been used in applications. Sankaran (1970) did not provide an ML estimate for the parameter p because of the computational difficulty to do so. However, the moment estimate is given by

$$\hat{p} = \frac{-(\bar{x}-1) + \sqrt{(\bar{x}-1)^2 + 8\bar{x}}}{2\bar{x}}$$

where, as usual, $\bar{x}$ denotes the sample mean.

It is interesting to note that the Lindlay distribution can be represented as a mixture of a Gamma and an exponential distributions, thus:

$$\pi \ \ Gamma(2, p) + (1-\pi) \ \ Exponential(p)$$

where $\pi = \frac{p}{p+1}$. This implies that the Poisson-Lindlay distribution can be considered to be a specific mixture of a negative binomial distribution with a geometric distribution.

The Poisson-Lindlay distribution is an one-parameter mixed Poisson distribution which can take a lot of shapes. For specific choices of p, the distribution can have a very long right tail.

2.3.3 The Poisson - Linear Exponential Family Distribution

Sankaran (1969) discussed the linear exponential family of continuous distributions as the mixing distribution of a mixed Poisson distribution. The linear exponential family contains distributions with probability density function

$$g(\theta) = \beta(p)\exp(-\theta p)h(\theta)$$

with $h(\theta) \geq 0$ and depending only on è and with p ranging over an interval on the real line, so that â(p) is finite and differentiable as many times as required, where $\frac{1}{\beta(p)} = \int \exp(\theta p)h(\theta)d\theta$.

Then, the resulting mixed Poisson distribution has probability function given by

$$P(x) = \frac{\beta(p)}{x!\beta(p+1)}\mu_x(p+1),$$

where $ì_x(p)$ denotes the x-th moment about the origin of the linear exponential family with parameter *p*. This representation requires the existence of the moments of all orders (which is equivalent to the differentiability of *â(p)*). Sankaran (1969) showed that for this family of distributions the moment estimates are close to the ML estimates and thus they can be used since they are easily obtainable.

The Poisson-Lindlay distribution is a member of this family and so is the geometric distribution. Another member of this one parameter family is the distribution considered by Kling and Goovaerts (1993). They used the name Linear exponential distribution, for the distribution with probability density function given by

$$g(\theta) = \frac{p[(p+1)\theta + 1]}{p+2} e^{-\theta p} \quad .$$

The resulting mixed Poisson distribution has probability function given by

$$P(x) = \frac{p^2(x+2)}{(p+1)(p+2)} \left(\frac{1}{p+1} \right)^x \quad .$$

A simple recursive form for the probabilities is

$$P(x+1) = \frac{(x+3)}{(p+1)(x+2)} P(x) \qquad \text{with} \qquad P(0) = \frac{2p^2}{(p+1)(p+2)} .$$

This distribution can also be written as a specific mixture of a negative binomial and a geometric distribution since the linear exponential distribution is of the form

$$\pi \ \ Gamma(2, p) + (1-\pi) \ \ Exponential(p)$$

with $\pi = \frac{p+1}{p+2}$. Comparing this distribution to the Lindlay distribution one may observe that they both are mixtures of the same distribution but with different mixing proportions.

The generalisation to a 2- parameter family of distributions is obvious, by considering that the mixing distribution is of the form

$$g(\theta) = \beta(p, a) \exp(-\theta p) h(\theta) .$$

This family of distributions contains the gamma distribution as well as many other distributions.

2.3.4 The Poisson-Lognormal Distribution

The lognormal distribution can also serve as a mixing distribution. The lognormal has more skewness than the Gamma distribution and thus it is more adequate for skew populations. Note that there are not many skewed distributions which can serve as mixing distributions and thus the lognormal distribution can fill this gap. So, if we assume that

$$g(\theta) = \frac{1}{\theta\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln\theta - m)^2}{2\sigma^2} \right),$$

the mixed Poisson-lognormal distribution has probability function given by

$$P(x) = \frac{1}{x!\sigma\sqrt{2\pi}} \int_0^\infty \theta^{x-1} \exp\left(-\frac{(\ln\theta - m)^2}{2\sigma^2} - \theta \right) d\theta \quad . \tag{2.25}$$

Unfortunately, the above probability function cannot be written in a closed form and this is the main reason for the limited use of this distribution. The numerical calculation for the probabilities and the inaccuracy of such approximation prohibits the wide use of the Poisson-lognormal distribution. Bulmer (1974) gave the following approximation for values of x≥10

$$P(x) \cong \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - m)^2}{2\sigma^2} \right) \left[1 + \frac{1}{2x\sigma^2} \left\{ \frac{(\ln x - m)^2}{\sigma^2} + \ln x - m - 1 \right\} \right] .$$

This was based on a Taylor expansion of second order of an expectation of a Gamma variate. Such approximations work satisfactorily for species abundance models where the Poisson-lognormal has found interesting applications (see, e.g., Cassie, 1964, Kempton and Taylor, 1974, among others). For count data, likely to be met in practise, the approximations are not sufficient and this makes the use of this distribution problematic.

Brown and Holgate (1971) presented tables of the Poisson-lognormal distribution for selected values of the parameters.

Cassie (1964) gave an interesting model for the Poisson-lognormal distribution. He supposed that some factors or covariates affect the values of the parameter of the simple Poisson distribution. So, if the parameter è can be modelled as $\theta = \exp(a_1 X_1 + a_2 X_2 + \ldots + a_k X_k)$ where X_i, i=1, 2,. . . , k follows a normal distribution, è follows a lognormal distribution giving rise to the Poisson-lognormal distribution. Cassie tried to describe the distribution of the number of insects in an area via a mixed Poisson distribution. He assumed that the Poisson parameter can be modelled using the exponential model described with covariates the temperature and the existence or not of some substances and other.

The application of the Poisson-lognormal distribution can be based on the moment estimates of the parameters. Even though the probability function of the Poisson-lognormal distribution is not available its moments can be easily obtained from the moments of the lognormal distribution.

2.3.5 Poisson-Confluent Hypergeometric Series Distribution

Bhatacharya (1966) proposed a family of distributions where the probability density function is expressed through the confluent hypergeometric function. A hypergeometric series is defined in general as

$$ {}_qF_p(a_1,a_2,\ldots,a_q;b_1,b_2,\ldots,b_p;z)=\sum_{r=0}^{\infty}\frac{a_1^{(r)}a_2^{(r)}\ldots a_q^{(r)}}{b_1^{(r)}b_2^{(r)}\ldots b_p^{(r)}}\frac{z^r}{r!} \quad , \qquad (2.26) $$

where $a^{(r)}=a(a+1)\ldots(a+r-1)=\dfrac{\Gamma(a+r)}{\Gamma(a)}$.

Depending on the values of p and q we may define several hypergeometric series. If we suppose that the mixing distribution is given by

$$ g(\theta)=\frac{b^a(b+1)^{p-1}}{\Gamma(p)}\theta^{p-1}\exp\left[-\theta(1+a)\right]M(a,b,\theta) \quad , \; a,b,\theta>0 \; , $$

where the $M(a,b,c)={}_1F_1(a,b,c)$ is called as the confluent hypergeometric function of the first kind, then the resulting mixed Poisson distribution has probability function given by

$$ P(x)=\frac{b^a(b+1)^{p-1}}{\Gamma(p)x!}\frac{\Gamma(x+p)}{(a+2)^{p+x}}{}_2F_1(a,p+x,p,1/(b+2)) . $$

For á=p the negative binomial is obtained. Because of the complicated form of such distributions the applicability is limited.

Recently, Ong (1996) derived a member of this family of distributions which is also the convolution of a negative binomial and a compound generalised hypergeometric factorial moment distribution.

2.3.6 The Poisson-Generalised Inverse Gaussian Distribution

Another mixed Poisson distribution which has been used in a variety of applications is the Poisson-generalised inverse Gaussian distribution. This distribution arises if the mixing distribution is the generalised inverse Gaussian distribution with parameters á, b and ã, denoted as *GIGD (á, b, ã)*. The generalised inverse Gaussian distribution has probability density function given by

$$g(\theta)=\frac{(a/b)^{\gamma/2}\theta^{\gamma-1}}{2K_{\gamma}\left(\sqrt{ab}\right)}\exp\left[-\frac{b/\theta+a\theta}{2}\right] \quad ,$$

where $K_x(a)$ is the modified Bessel function of order x. The domain of the parameters is more complicated. In general, $\gamma\in\Re$, and á,b >0. However for positive ã, b may be 0, and for negative ã , á is allowed to be 0. Then, the resulting Poisson-generalised inverse Gaussian distribution has probability function given by

$$P(x)=\left(\frac{a}{a+2}\right)^{\gamma/2}\left(\frac{b}{a+2}\right)^{x/2}\frac{K_{\gamma+x}\left(\sqrt{(a+2)b}\right)}{x!K_{\gamma}\left(\sqrt{ab}\right)} \quad . \qquad (2.27)$$

This distribution was first considered by Good (1953). Sichel (1974,1975,1982) gave a more detailed description of this distribution, and Atkinson and Yeh (1982) and Stein *et al.* (1987) described its properties as well as multivariate analogues and ML estimates for the parameters.

The existence of the modified Bessel function in the probability function makes the direct evaluation of probabilities difficult. However using the following recursive form the probabilities are easily obtainable:

$$bP(x-1)=x(x+1)(a+2)P(x+1)-2x(\gamma+x)P(x) \text{ for } \quad x\geq 1$$

This recursion is unstable and it has to be used with caution. *P(0)* and *P(1)* have to be calculated numerically.

The generalised inverse Gaussian distribution is a conjugate distribution for the Poisson parameter. So, if the prior distribution is a *GIGD (a,b,ã)* distribution the posterior distribution of *è|x* is a *GIGD (a+2,b,ã+x)* distribution.

This distribution contains as a special case the negative binomial distribution for *b=0*. Another very interesting special form is the inverse Gaussian distribution which is obtained when *ã=-1/2*. In this case the resulting Poisson-inverse Gaussian distribution is much simplified and it will be discussed in the sequel. Bivariate extensions are described in Kocherlakota and Kocherlakota (1992). Another limiting case of the generalised inverse Gaussian distribution distribution is the inverse Gamma distribution which stems from the generalised inverse Gaussian distribution when ã<0 and á=0. The use of Poisson- generalised inverse Gaussian distribution distribution is limited mainly because of its complicated form. The Poisson-inverse Gaussian, however, is a widely used member of the family of mixed Poisson distributions.

2.3.7 The Poisson-Inverse Gaussian Distribution

If we assume that the distribution of è is the inverse Gaussian distribution, i.e. its probability density function is given by:

$$g(\theta) = \sqrt{\frac{\sigma}{2\pi\theta^3}} \exp\left(\frac{-\sigma(\theta-\mu)^2}{2\mu^2\theta}\right) \qquad \text{ì, ó, è > 0,}$$

then X follows the Poisson-inverse Gaussian distribution. Its probability function is given by (2.27) for *ã=-1/2*. This form is very complicated, but, fortunately, simple recurrence relations exist for the calculation of the probabilities. So we can calculate the probabilities using the following iterative scheme:

$$P(0) = \exp\left[\frac{\mu}{\beta}\left(1-\sqrt{1+2\beta}\right)\right], \qquad P(1) = \frac{\mu}{\sqrt{1+2\beta}}P(0) \qquad \text{and}$$

$$P(x) = \frac{2\beta}{1+2\beta}\left(1-\frac{3}{2x}\right)P(x-1) + \frac{\mu^2}{x(x-1)(1+2\beta)}P(x-2) \qquad \text{for x=2, 3, ...}$$

where the parameter â relates to the parameters of the mixing distribution through the formula â=ì^2 /ó. The mean and the variance are $E(X)=\mu$ and $Var(X)=\mu(1+\beta)$. This distribution is infinitely divisible and hence it can also be obtained as a compound Poisson distribution, when the distribution of the X_i's is the extended negative binomial (see, Engen, 1974) with probability function given by

$$P(x) = \frac{\Gamma(x-0.5)\left(\frac{2b}{1+2b}\right)^n}{2x!\,\Gamma(0.5)\left(1-\frac{1}{\sqrt{1+2b}}\right)} \qquad , \qquad \text{x=1,2, ...}$$

Sichel (1975) showed that the sum of i.i.d. Poisson-inverse Gaussian random variates is again a Poisson-inverse Gaussian random variate, due to the reproductive property of the inverse Gaussian distribution. (See Folks and Chikara, 1978). Ord and Whitmore (1986) and Willmot (1987) introduced this distribution in the context of species abundance models and actuarial applications respectively. Dean *et al.* (1989) described a Poisson-inverse Gaussian regression model.

Kaas and Hesselager (1995) tried to compare three mixed Poisson distributions, namely the negative binomial , the Poisson-lognormal and the Poisson-inverse Gaussian

distribution. The comparison is made on the mixing distribution since as already seen the behaviour of the mixing density g is strongly related to the behaviour of the *MP(g)*. They showed that the Poisson-lognormal has heavier and longer tails, as well as more skewness and kurtosis for the case of the same mean and variance. Unfortunately, more formal comparisons of a larger number of mixed Poisson distributions are not known and it remains an open problem to try to compare their behaviour.

2.3.8 The Poisson- Inverse Gamma distribution

If we assume that the mixing distribution is a inverse Gamma distribution with probability density function given by

$$g(\theta) = \frac{b^a e^{-b/\theta}}{\Gamma(a)\theta^{a+1}} \quad , \tag{2.28}$$

the resulting Poisson-Inverse Gamma distribution has probability function given by

$$P(x) = \frac{K_{x-a}\left(2\sqrt{b}\right)}{x!\Gamma(a)} 2b^{(x+a)/2}$$

(Willmot, 1993). The recursive formula for evaluating the probabilities is given by:

$$(x+1)(x+2)P(x+2) = (x+1)(x+1-a)P(x+1) + bP(x)\,.$$

The numerical calculation of the first two probabilities is needed in order to obtain the probability function of this distribution. Applications of this distribution are not known. Note that the Poisson-inverse Gamma is a special case of the Poisson-generalised inverse Gaussian described in the previous section.

2.3.9 The Poisson - Truncated Normal Distribution

The normal distribution itself is not a plausible choice as the mixing distribution because of its support in the negative axis. The truncated normal distribution, however, can be considered as a mixing distribution. In this case the probability density function of è is given by

$$g(\theta) = \frac{1}{\Phi(\mu/\sigma)\sigma\sqrt{2\pi}} \exp\left(-\frac{(\theta-\mu)^2}{2\sigma^2}\right) \qquad \text{è}\geq 0,\ \text{ì,ó} > 0,$$

where $\Phi(x)$ is the cumulative distribution function at x of the standard normal distribution. Note that ì and ó are the mean and the standard deviation of the untruncated normal. The probability function of the resulting Poisson-truncated normal distribution cannot be written in a closed form. However, applying the method discussed in Willmot (1993) a recursive relation can be found. So, we can calculate it by:

$$P(0)=\frac{\Phi\left(\frac{\mu-\sigma^2}{\sigma}\right)}{\Phi\left(\frac{\mu^2}{\sigma}\right)}\exp(-\mu+\sigma^2/2)\,, \qquad P(1)=(\mu-\sigma^2)P(0)+\sigma^2\frac{\exp(-\mu^2/2\sigma^2)}{\sigma\sqrt{2\pi}\Phi(-\mu/\sigma)}$$

$$P(m+1)=\frac{\sigma^2\big(P(m-1)-P(m)\big)+\mu P(m)}{(m+1)} \qquad \text{for m=2,3,. . .}$$

The distribution was defined in Patil (1964. The shape of this distribution resembles well the bell-shape of the normal distribution.

The Hermite distribution, examined in Kemp and Kemp (1965) can be regarded as a compound Poisson distribution, with a binomial distribution with n=2 as the summand distribution. Formally, the distribution can also be regarded as a mixed Poisson, with mixing distribution the normal distribution. This result lacks physical interpretation since the parameter of the Poisson distribution has to be positive. However, we may assume that if the parameters of the normal distribution give very small probability at the negative axis, the normal distribution and the normal distribution truncated at 0 will be almost identical and hence the Hermite distribution in this case is very similar to the Poisson-truncated normal. The Hermite distribution has probability function given by

$$P(x)=\exp\big(-(a_1+a_2)\big)\sum_{j=0}^{[x/2]}\frac{{a_1}^{x-2j}{a_2}^{j}}{(x-2j)!\,j!} \qquad \text{, x=0, 1, ..., á}_1 \text{ , á}_2>0\text{ ,}$$

where, $á_1= 2ì- ó^2$ *and* $á_2 = (ó^2 -ì)/2$, and [a] is the integer part of a. The probabilities can be easily calculated via the following iterative scheme:

$P(0)= exp\{-(á_1 + á_2)\}$, $P(1)= P(0)\, á_1$ and

$$P(x+1)=\frac{a_1P(x)+2a_2P(x-1)}{x+1} \qquad \text{, for x>1} \qquad .$$

Multivariate extensions of the Hermite distribution can be found in Kocherlakota and Kocherlakota (1992).

2.3.10 The Generalised Waring Distribution

Another interesting member of the family of mixed Poisson distributions is the generalised Waring distribution (GWD) proposed by Irwin (1968) to describe the frequency distribution of the number of accidents. It can describe data sets with a long right tail, and it has a larger tail than the negative binomial from which can be obtained via mixing. The probability function of the generalised Waring distribution is

$$P(x)=\frac{\Gamma(a+c)\Gamma(b+c)\Gamma(a+x)\Gamma(b+x)}{\Gamma(a)\Gamma(b)\Gamma(c)\Gamma(a+b+c+x)x!}\text{, x=0,1,..., á,b,c>0} \quad . \quad (2.29)$$

A simple recursive scheme is available for efficient calculation of the probabilities. This scheme is

$$P(x+1)=\frac{(a+x)(b+x)}{(x+1)(a+b+c+x)}P(x) \quad \text{, x=0,1,...}$$

Only *P(0)* is needed to be evaluated numerically. To do so we need to evaluate the Gamma functions.

Irwin (1968) derived the generalised Waring distribution in the context of accident theory by starting from a Poisson distribution. The parameter ë of the Poisson distribution follows itself a Gamma distribution with parameters p and b representing thus the difference in proneness of different individuals. This results in a negative binomial distribution with parameters b and p, given in (2.22). He allowed the parameter p of the negative binomial to follow a Beta type II distribution with parameters á and c, given in (1.7), representing in this way the difference in the liability of different individuals. Then he obtained the generalised Waring distribution. This derivation can be represented as

$$Poisson(\lambda)\underset{\lambda}{\wedge}Gamma(p,k)\underset{p}{\wedge}BetaII(a,c)$$

Since for a random variable X following the *BetaI(á,b)* distribution the random variable X/(1+X) follows the *BetaII(á,b)* distribution the generalised Waring distribution can be represented as

$$Poisson(\lambda)\underset{\lambda}{\wedge}Gamma(p,k)\underset{p/(1-p)}{\wedge}BetaI(a,c)$$

The generalised Waring distribution is also a mixed negative binomial distribution of the form

$$NegativeBinomial(b,p) \underset{p}{\Lambda} BetaII(a,c)$$

or equivalently

$$NegativeBinomial(b,p) \underset{p/(1-p)}{\Lambda} BetaI(a,c)$$

This stems from the associative property of mixtures given in section 1.2. It is worth noting that for the negative binomial distribution defined in (2.22) the Beta type II distribution is conjugate for the parameter á. In other words if the prior is a *BetaII (á,c)* then the posterior is *BetaII(a+b,c+x).*

The assumptions of Irwin separated the notion of accident proneness in two components. The first refers to the person's predisposition to accidents and it is called proneness and the second refers to external factors which make an individual be more probable to incur an accident and it is called liability. Such representation enables one to separate the total variance of the data in three factors: randomness, proneness and liability. Unfortunately, the symmetry of the generalised Waring distribution about the parameters á and b, makes our assumptions for the model of limited practical use. Thus we can separate the total variance in three components but we cannot assign the proportion of variance to proneness or to the liability (see, e.g. Xekalaki, 1983a).

Xekalaki (1984a,b), in order to overcome this difficulty, defined a bivariate generalised Waring distribution and used data from two successive periods. Under such a model it is possible to distinguish between proneness and liability, based on the initial assumptions for the model.

Irwin (1975) examined in depth some of the properties of the generalised Waring distribution. The infinite divisibility of the generalised Waring distribution was shown in Xekalaki (1983b). Estimation techniques are discussed in Xekalaki (1987). Thus the generalised Waring distribution can be represented as a compound Poisson distribution. The discrete distribution giving rise to the generalised Waring distribution as a compound Poisson is not known in closed form. Its probabilities can be obtained via the recursive scheme given in (2.8).

A more formal derivation of the generalised Waring distribution as a mixed Poisson distribution requires the mixing distribution to be of the form:

$$g(\theta) = \frac{\Gamma(b+c)}{\Gamma(a)\Gamma(b)\Gamma(c)} \theta^{a-1} \int_0^{\infty} e^{-\theta t} \frac{t^{a+c-1}}{(1+t)^{b+c}} dt =$$

$$= \frac{\Gamma(b+c)\Gamma(a+c)}{\Gamma(a)\Gamma(b)\Gamma(c)} \theta^{a-1} \Psi(a+c, b-a-1, \theta) \quad ,$$

where *Ø(a,c,x)* is the confluent hypergeometric series of the second kind. This distribution is known as gamma product ratio distribution (Sibuya, 1979) and can be obtained by mixing a Gamma distribution with a Beta type II distribution. More formally as

$$Gamma(p,k) \underset{p}{\Lambda} BetaII(a,c) .$$

Its name is based on the fact that if X_i follows a $Gamma(1, m_i)$ distribution, i=1,2,3, and $m_1 = a$, $m_2 = b$ and $m_3 = c$, then the distribution of the random variable $Y = \frac{X_1 X_2}{X_3}$ follows the Gamma product ratio distribution. Devroye (1993) used this representation for simulating random variates from the generalised Waring distribution.

The r-th factorial moment is given by

$$E\left[X(X-1)...(X-r+1)\right] = \frac{\Gamma(a+1)\Gamma(b+1)\Gamma(c-r)}{\Gamma(a-r+1)\Gamma(b-r+1)\Gamma(c)} \quad , \text{ c>r.}$$

Some special forms of the generalised Waring distribution are the simple Waring distribution, discussed in an ecological application by Pielou (1962) and in a biometrical application by Weinberg and Gladen (1986). The simple Waring distribution results from the generalised Waring distribution when b=1 and has probability function of the form

$$P(x) = \frac{c\Gamma(a+c)\Gamma(a+x)}{\Gamma(a)\Gamma(a+c+x+1)} \quad .$$

Pielou (1962) and Weinberg and Gladen (1986) derived the Waring distribution by mixing the Geometric distribution given in (2.24) with a Beta type II distribution.

Another special form of the generalised Waring distribution (and the Waring distribution) is the Yule distribution obtained from the generalised Waring distribution when b=á=1. Its probability function is given by

$$P(x) = \frac{c\Gamma(c+1)x!}{\Gamma(c+x+2)} \quad .$$

The Yule distribution is a one-parameter distribution with very long right tail and it has been used for describing the income distribution (see Simon, 1955, Xekalaki, 1984c, Panaretos, 1989).

2.3.11 The Poisson -Beta Distribution

The Beta distribution (in a variety of forms) can also be regarded as a mixing distribution. Kempton (1975), in a species abundance model, considered a generalised Beta distribution as the mixing distribution. The distribution used in Kempton (1975) under the name full Beta distribution is the so-called generalised Pareto distribution (Willmot, 1993) with probability density function given by

$$g(\theta)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\mu^a\theta^{b-1}}{(\mu+\theta)^{a+b}} \quad ,\theta,a,b,\mu>0 \qquad . \qquad (2.30)$$

then the resulting Poisson-generalised Pareto distribution has probability function given by

$$P(x)=\frac{\mu^x\Gamma(a+b)\Gamma(x+b)}{\Gamma(b)\Gamma(a)x!}\Psi(b+x,x+1-a,\mu) \qquad .$$

Note that for μ=1 in (2.30) the Beta Type II distribution is obtained. For b=1 the simple Pareto distribution is obtained. We will present the Poisson-Pareto distribution in the sequel. The generalised Pareto distribution can be obtained as a Gamma mixture of the form

$$Gamma(p,b)\underset{p}{\Lambda}Gamma(\mu,a)$$

The recurrence relations for evaluating the probabilities are (see Willmot, 1993):

$$(x+2)(x+1)P(x+2)=(x+1)(x+1-a-\mu)P(x+1)+\mu(x+b)P(x)\,. \qquad (2.31)$$

The direct evaluation of the first two probabilities is necessary. Ong and Muthaloo (1995) used a slightly different parameterisation.

The r-th factorial moment $\mu_{(r)}$ of this distribution is:

$$\mu_{(r)}=\frac{\Gamma(b+r)\Gamma(a-r)}{\Gamma(b)\Gamma(a)}\mu^r \qquad , \qquad \alpha\geq r \qquad .$$

Using the associative property of mixtures the Poisson-generalised Pareto distribution is also a mixed negative binomial distribution of the form

$$NegativeBinomial(p,a)\underset{a}{\Lambda}Gamma(b,q)\,.$$

Kempton (1975) showed that for some values of the parameters this distribution tends to the negative binomial distribution. The distribution differs from the generalised Waring distribution in that a Gamma distribution is considered for the parameter of the negative

binomial rather than a Beta distribution. Since the Beta distribution has a longer right tail than the Gamma distribution this distribution has shorter tails than the generalised Waring distribution.

A special case of the full Beta model is the Beta type II distribution which results from the full Beta distribution by setting μ=1. Holla and Bhattacharya (1965) described this Poisson-Beta type II distribution as well as multivariate analogues.

For the Poisson-BetaII distribution, the recursive scheme given in (2.31) may be used with μ=1. Numerical evaluation of the first probabilities is needed. Kempton (1975) also discussed the case when p=0 in the full beta distribution. This always results in a J-shaped distribution which tends to the logarithmic series distribution. Willmot (1986) discussed a mixed Poisson distribution with a mixing distribution of the form

$$g(\theta) = \frac{\theta^{a-1} b(\mu-\theta)^{b-1}}{B(a,b)\mu^{a+b-1}} \qquad , \qquad 0 < \text{è} < \text{ì}$$

for ì=1 we obtain the Beta Type I distribution. Such a mixing distribution restricts the range of the parameter of the Poisson distribution. Gurland (1958) first described the Poisson-Beta type I distribution. (See also Holla and Bhattacharya, 1965). The resulting mixed Poisson distribution has probability function given by

$$P(x) = \frac{\mu^x B(a+x,b)}{B(p,q)x!} M\left(a+x, a+x+b, -\mu\right) \qquad , \qquad (2.32)$$

where *M(a,b,c)* is the confluent hypergeometric series of the first kind. The recursive formula for the calculation for the probabilities is

$$x(x+1)P(x+1) = x\left(x-1+a+b+\mu\right)P(x) - \mu\left(x+a-1\right)P(x-1)$$

which is unstable. Gurland (1958) suggested the calculation of the probabilities from their probability function, using the infinite series representation of the hypergeometric function, evaluated with a large number of summands.

Beall and Rescia (1953) proposed the above distribution with á=1, which is a much simpler form. (See also Willmot, 1986).

2.3.12 The Poisson-Uniform Distribution

Another mixed Poisson distribution is the Poisson-uniform distribution, which can be obtained from (2.32) for á=b=1. This has as mixing distribution a uniform distribution in the

interval [0,ì]. A more general uniform mixing distribution was considered by Bhattacharya (1967). He used the uniform distribution in the interval [á,b] i.e.

$$g(\theta)=\frac{1}{b-a} \qquad , \qquad 0<\text{á}<\text{è}<\text{b} \qquad .$$

The resulting distribution has probability function of given by

$$P(x)=\frac{1}{x!(b-a)}\int_a^b \theta^x e^{-\theta}d\theta$$

The integral may be represented using the incomplete gamma function. However, a simple recursive formula is available which reduces the effort for calculating the probabilities. Thus, we obtain

$$P(x+1)=P(x)+\frac{e^{-a}a^{x+1}-e^{-b}b^{x+1}}{(x+1)!} \quad \text{with } P(0)=\frac{1}{b-a}\left(e^{-a}-e^{-b}\right)$$

It is interesting that in the above recursive formula the difference between two Poisson distribution with parameters á and b is added at each step. In spite of its simplicity this distribution has not been used in applications, mainly because the uniform assumption does not seem plausible. Bhattacharya (1967) showed that if the mixing distribution is a uniform distribution in the interval [a,b], the resulting posterior distribution is a Gamma distribution truncated at the points a and b from the left and from the right respectively. This implication advises the use of the Gamma distribution as the mixing distribution by assuming that in the start of their lives individuals possess a uniform distribution but as time passes the distribution describing the inhomogeneity of the population is a gamma distribution. (Recall that the Gamma distribution is conjugate for the Poisson distribution) .

Assuming as a mixing distribution the right-truncated Gamma at point b, with probability density function:

$$g(\theta)=\frac{\mu^a\theta^{a-1}\exp(-\theta\mu)}{\Gamma(a)\{1-\Gamma(a,\mu b)\}} \quad , \text{è}>\text{b} \quad ,$$

we obtain the Poisson-truncated Gamma distribution with probability function given by

$$P(x)=\frac{\mu^a\Gamma(x+a)}{(1+\mu)^{a+x}\Gamma(a)}\left\{1-\Gamma(a+x,(1+\mu)b\right\}.$$

Willmot (1993) gave the following recursive formula for the probabilities

$$(\mu+1)(x+1)P(x+1)=(x+a)P(x)+\Gamma(b)\frac{b^{x+1}e^{-b}}{(x+1)!} \quad .$$

2.3.13 The Poisson-Modified Bessel Function of the Third Kind Distribution

Ong and Muthaloo (1995) described another mixed Poisson distribution whose probability function is expressed through the confluent hypergeometric function. In this distribution the mixing distribution is the mixture of a *Gamma (a,p)* distribution when a follows an inverse gamma distribution with probability density function defined in (2.28). This mixing distribution is known as the modified Bessel function of the third kind because of the presence of the relevant modified Bessel function in its probability density function. This distribution can be formally written as

$$Gamma(p,a) \underset{p}{\Lambda} InverseGamma(b,c)$$

and its form as a mixed Poisson distribution is

$$Poisson(\lambda) \underset{\lambda}{\Lambda} Gamma(p,a) \underset{p}{\Lambda} InverseGamma(b,c) \ .$$

This distribution is also a mixed negative binomial distribution, with the inverse Gamma distribution as mixing distribution, i.e. it can be represented as

$$NegativeBinomial(p,a) \underset{a}{\Lambda} InverseGamma(b,c) .$$

Then the resulting mixed Poisson distribution has probability function of the form

$$P(x) = \frac{c^x \Gamma(a+x)}{xB(x,b)\Gamma(a)} \Psi(a+x, a+1-b, 1) , \qquad (2.33)$$

The useful recurrence relation for the probabilities of this distribution is:

$$x(x+1)P(x+1) = x\left(2x-1+a+b+1/c\right)P(x) - \left(b+x-1\right)\left(x+a-1\right)P(x-1) .$$

The r-th factorial moment $\grave{\imath}_{(r)}$ of this distribution is

$$\mu_{(r)} = \frac{\Gamma(a+1)}{\Gamma(a+r-1)} \frac{\Gamma(b+1)}{\Gamma(b+r-1)} c^r \qquad .$$

Ong and Muthaloo (1995) showed that this distribution is very flexible for describing long tailed data.

We would like to point out the interesting resemblance between three members of the family of the mixed Poisson distributions, namely between the generalised Waring distribution given in (2.29), the Poisson-generalised Pareto given in (2.30) , and the Poisson-modified Bessel function of the third kind distribution given in (2.33). The connection is that all these distributions arise from mixing distributions which are mixed Gamma distributions

with a Beta, a Gamma and an inverse Gamma mixing distributions, respectively. This similarity can help us treat all these three distributions in a unified manner. For example, the property of decomposition of variance, known for the generalised Waring distribution can be also achieved for the other two distributions. These distributions are also mixed negative binomial distributions.

2.3.14 Dellaporte Distribution

The Dellaporte distribution is the distribution of the convolution of a Poisson distribution with a negative binomial distribution. Ruohonen (1988) showed that it is a mixed Poisson distribution with mixing distribution a three parameter gamma distribution, namely a shifted Gamma distribution with probability density function given by

$$g(\theta) = \frac{a^b}{\Gamma(b)}(\theta - c)^{b-1} \exp\left[-a(\theta - c)\right], \qquad \text{è>c.}$$

Note that for c=0 we obtain the Gamma distribution. The probability function of the resulting Dellaporte distribution is given by:

$$P(x) = \sum_{n=0}^{x} \frac{c^{x-n}e^{-c}}{(x-n)!} \frac{\Gamma(n+b)}{n!\Gamma(b)} \left(\frac{1}{1+a}\right)^n \left(\frac{a}{1+a}\right)^b \quad . \tag{2.34}$$

It is recognisable that this is the convolution of Poisson distribution with parameter c distribution with a negative binomial distribution with parameters b and a. This representation is useful for obtaining the moments of the distribution. However, the calculation of the probabilities directly from the probability function is computationally intensive and thus a recursive scheme proposed by Willmot (1993) is more useful. Under this scheme the probabilities may be calculated using the relation

$$(a+1)(x+1)P(x+1) = \left[b + c(a+1) + x\right]P(x) - cP(x-1) \qquad \text{for x≥1.}$$

The first two probabilities *P(0)* and *P(1)* can be calculated as:

$$P(0) = e^{-c}\left(\frac{a}{1+a}\right)^b \text{ and } P(1) = P(0)\left(c + \frac{b}{1+a}\right).$$

Willmot and Sundt (1989) proposed the use of the Dellaporte distribution in an actuarial context giving useful formulas based on the convolution representation. Ruohonen (1988) derived estimates for the parameters using both the moment method and the ML method as well as another method which uses the zero frequency. He also applied the Dellaporte distribution to actuarial data, and he derived some interesting properties of the distribution.

2.3.15 The Family of Poisson - Pareto Distributions

A special case of the Poisson-generalised Pareto distribution given in (2.30) is the Poisson-Pareto distribution, arising form the Poisson-generalised Pareto if b=1. The recurrence relations for evaluating the probabilities are (see Willmot, 1993):

$$(x+2)(x+1)P(x+2) = (x+1)(x+1-a-\mu)P(x+1) + \mu(x+b)P(x) \qquad .$$

The direct evaluation of the first two probabilities is necessary. However for the case of the Poisson-Pareto distribution we have also the formula:

$$P(1) = a - (a+\mu)P(0)$$

which requires the numerical evaluation of P(0) only.

Another Pareto distribution is the shifted Pareto with probability density function

$$g(\theta) = \frac{a}{\mu}\left(\frac{\mu}{\theta}\right)^{a+1} \qquad , \text{è} > \text{ì} \quad .$$

Since this distribution is a shifted at ì version of the simple Pareto distribution, the resulting mixed Poisson-shifted Pareto distribution is the convolution of a Poisson distribution and a Poisson-Pareto distribution. The probability function is given by

$$P(x) = \frac{a\mu\Gamma(\mu-a)}{x!}\{1-\Gamma(x-a,\mu)\},$$

where $\tilde{A}(\acute{a},x)$ is the incomplete Gamma function. The probabilities can be obtained either directly from the probability function, evaluating numerically the incomplete Gamma function or via the following recursive scheme:

$$(x+1)P(x+1) = (x-a)P(x) + a\frac{\mu^x e^{-\mu}}{x!}.$$

2.3.16 Other Mixed Poisson Distributions

Apart from the variety of mixed Poisson distributions described above there are several other members of this family which have not been studied at all. From the above mentioned distributions very few have been studied in depth, and for the rest only a few references are known. The applicability of these distributions depends on the tractability of their probability function. However, the recursive relations which are available via the fundamental method of Willmot (1993) facilitate the calculation of the probability function. We will mention some other distributions which have appeared in the literature, mainly as references and not as complete examples.

In this category belong the two mixed Poisson distributions used by Burrel and Cane (1982) in the analysis of library data. These two distributions were derived from a well specified model in library data analysis. Rai (1971) proposed the use of the power function distribution with probability density function

$$g(\lambda) = \frac{k}{\theta^k}\lambda^{k-1} \qquad , \qquad 0 < \lambda < \theta$$

as the mixing distribution. For k=1 the Poisson-uniform distribution is obtained while for k=2 a Poisson-triangular distribution is obtained. Willmot (1993) discussed power transformations of standard distributions like the Gamma distribution and the Beta distribution, namely the distribution of X^c where X is Gamma or Beta distributed. Some known distributions belong to this category such as the Weibull, the Burr distribution etc. Willmot (1986) showed that the noncentral chi-square distribution is the convolution of a Gamma distribution with a $Poisson(m/2) \underset{m}{\wedge} Gamma(1,2)$ distribution. Hence the mixed Poisson-noncentral chi-square distribution can be defined as the convolution of a negative binomial with the $Poisson(\lambda) \underset{\lambda}{\wedge} \left[Poisson(m/2) \underset{m}{\wedge} Gamma(1,2) \right]$ distribution. Note that this is a convolution of two negative binomial distributions with varying parameter p. It is known that the convolution of two negative binomial distributions with the same parameter p is again a negative binomial distribution.

Willmot (1986) proposed the use of the reciprocal inverse Gaussian distribution which is the convolution of a $Gamma\left(\frac{1}{2}, \frac{2}{m}\right)$ distribution with an inverse Gaussian distribution with

parameters *ì* and *1/m*. Thus the resulting Poisson-reciprocal inverse Gaussian distribution is the convolution of a negative binomial and a Poisson- inverse Gaussian distribution.

Willmot (1993) also discussed the use of the exponential-inverse Gaussian distribution as the mixing distribution. Formally the distribution arises as

$$Exponential(\theta) \underset{\lambda}{\wedge} InverseGaussian(\mu, \sigma).$$

This distribution is also a mixed geometric distribution.

Philipson (1960) and Albrecht (1982) discussed in general the use of the members of the Pearson's family of continuous distributions as mixing distributions. Some special members of this family have been treated in detail such as the Gamma distribution, the Beta distribution etc. The importance in this presentation is that Albrecht (1982) proposed the use of moment estimates in order to choose between the members of this family, since the first four moments determine uniquely the members of the Pearson's family.

Few years later, Albrecht (1984) described several mixed Poisson distributions based on the Mellin and Laplace transforms of their mixing distributions. Interesting members of the distributions described by Albrecht are the F-distribution, the Maxwell, the Rayleigh distribution, the Weibull distribution, the chi-distribution and some members of the Pearson family. All these distributions have probability function which involve special functions. This results in reducing their applicability. For example, the Gamma function is regarded as much simpler than the parabolic cylinder function (see Abramowitz and Stegum, 1965) appearing in the probability function of the Poisson-Maxwell distribution.

Gaver and O'Muircheartaigh (1987) used the log-Student distribution in the context of empirical Bayes estimation of the parameter of the Poisson distribution. This is the distribution of a random variable $Y = e^X$ where X follows the Student distribution. For such a mixing distribution the mixed distribution is very complicated and numerical methods are necessary for the calculation of the probabilities.

All the mixed Poisson distributions discussed above have support on the nonnegative axis. In fact some discrete distributions have restricted support on the positive integers, namely x =1,2, ... The logarithmic series is the best known such example. The support of the Poisson distribution contains the value x=0. We may transform the Poisson distribution to the nonzero integers by considering either shifted, or truncated versions of the Poisson distribution. Shifted versions will lead to mixed shifted Poisson distributions. A Poisson distribution shifted at r has probability function of the form

$$P(x) = \frac{e^{-\lambda}\lambda^{(x-r)}}{(x-r)!} \qquad \text{x=r, r+1, ...}$$

If we mix this shifted Poisson, with a mixing density g the result will clearly be an *MP(g)* distribution shifted at r.

This is not true for the truncated Poisson distribution. A truncated at 0 Poisson distribution has probability function given by

$$P(x) = \frac{e^{-\lambda}\lambda^{-x}}{\left(1-e^{-\lambda}\right)x!}, \quad \text{x=1, 2,....,}$$

and hence the mixtures of this distribution cannot be related directly to the mixed Poisson distributions.

Some zero-truncated mixtures are also known in the literature. The logarithmic series distribution is a mixed-zero truncated Poisson distribution (see Sibuya, 1979) with mixing distribution having probability density function given by

$$g(\theta) = \frac{\left(1-e^{-\theta}\right)e^{-\theta/p}}{\theta \log(1+p)}, \qquad \theta, p > 0 \qquad .$$

The digamma and trigamma distributions discussed by Sibuya (1979) are also mixed-zero truncated Poisson distributions due to their derivation as mixed logarithmic series distributions.

In concluding, we provide Table 2.1 which summarises the main mixed Poisson distributions described in this chapter.

Table 2.1
Some mixed Poisson distributions

mixed Poisson distribution	**mixing distribution**	**a key reference**
Negative Binomial	Gamma	Greenwood and Yule (1920)
Geometric	Exponential	Johnson *et al.* (1992)
Poisson- Linear Exponential Family	Linear Exponential Family	Sankaran (1969)
Poisson-Lindlay	Lindlay	Sankaran (1970)
Poisson-Linear Exponential	Linear Exponential	Kling and Goovaerts (1993)
Poisson-Lognormal	Lognormal	Bulmer (1974)
Poisson-Confluent Hypergeometric Series	Confluent Hypergeometric Series	Bhattacharya (1966)
Poisson-generalised inverse Gaussian	Generalised Inverse Gaussian	Sichel (1974)
Sichel	Inverse Gaussian	Sichel (1975)
Poisson-Inverse Gamma	Inverse Gamma	Willmot (1993)
Poisson-Truncated Normal	Truncated normal	Patil (1965)
Generalised Waring	Gamma Product Ratio	Irwin (1975)
Simple Waring		Pielou (1962)
Yule		Simon (1955)
Poisson-Generalised Pareto	Generalised Pareto	Kempton (1975)
Poisson-Beta I	Beta Type I	Holla Bhattacharya (1965)
Poisson-Beta II	Beta Type II	Gurland (1957)
Poisson-Truncated BetaII	Truncated Beta type II	Willmot (1986)
Poisson -Uniform	Uniform	Bhattacharya (1966)
Poisson-Truncated Gamma	Truncated Gamma	Willmot (1993)
Dellaporte	Shifted Gamma	Ruohonen (1988)
Poisson-Modified Bessel of the 3rd kind	Modified Bessel of the 3rd kind	Ong and Muthaloo (1995)
Poisson-Pareto	Pareto	Willmot (1993)
Poisson-Shifted Pareto	Shifted Pareto	Willmot (1993)
Poisson-Pearson Family	Pearson's family of distributions	Albrecht (1982)
Poisson-Log-Student	Log-Student	Gaver and O'Muircheartaigh (1987)
Poisson-Power function	Power Function distribution	Rai (1971)

2.4 Discussion

A wide variety of mixed Poisson distributions was presented in this chapter with the aim of extending the use and the relations among the members of this large family of discrete distributions. The need for methods of estimation for many of them is obvious in order to increase their applicability. A complete examination of several members of this family has not been made, even though it would be also of interest.

We would also like to note another point, which reveals some interesting properties of mixed Poisson distributions as well as models leading to these mixed Poisson distributions. Consider the case when each individual follows a Po(ë) distribution. The parameter ë may be regarded as a random variable too. However, we may consider that two different factors affect the value of ë. To describe this relation we may assume either

- an additive model of the form ë=ì+í or
- a multiplicative model of the form ë=ìí

The factors ì and í can be regarded as random variables having probability density functions g_1 and g_2 respectively. A well known example from accident theory is the distinction of the factors contributing to an accident into proneness and liability, (see Xekalaki, 1983a). The notion can be extended to other fields, too.

As far as the first model is concerned, it is known that the resulting mixed Poisson distribution will have as a mixing density the convolution of g_1 and g_2. Hence mixing distributions with the reproductive property can be regarded as resulting from such a scheme. The Dellaporte distribution may arise from this additive model. Other interesting examples of this kind are given in Barndorrf-Nielsen *et al.* (1992) where the convolution of a Gamma with a generalised inverse Gaussian distribution leads to another generalised inverse Gaussian distribution. The model is rather conceptual as the parameters are confounded and not easily estimated if no additional information is available.

Let us now turn to the multiplicative model. If we use Proposition 1.1, we obtain that ë is a random variable which stems from a mixture model. Interesting examples of this type are the mixtures of the Gamma distribution regarded as mixing distributions giving rise to the generalised Waring distribution, the Poisson-generalised Pareto distribution and to the Poisson-Modified Bessel distribution. These distributions can be regarded as resulting from such a multiplicative model. Unfortunately, the parameters are not identifiable, i.e. one does not know which distribution corresponds to ì and which to í. This problem was also

mentioned when examining the generalised Waring distribution. However the model is very interesting itself.

Multiplicative Poisson models have been used in latent trait models (e.g. Jansen and Van Duijn ,1992). Such models assume that the test score depends on both the ability of the subject and the difficulty of the test which are both random variables. Assuming that they are related multiplicatively, the multiplicative Poisson model is used to describe such models. It is of interest and open for further research that models where both factors are themselves random variables do not appear to have been considered in the literature.

An interesting special case is when the distribution of ì is a degenerate distribution. This corresponds to the case where ì is no longer a random variable but it has a known value. This is the case where we have a *Po(ìë)* distribution and ë is a random variable. In general this distribution is different from the one obtained in the simple case ì=1. However, for mixing distributions with a scale parameter the resulting mixed Poisson distribution is of the same form. Such examples are the Gamma distribution and the inverse Gaussian distribution. A counterexample is the Beta distribution (of any type).

Finally Carriere (1993) described nonparametric tests applicable to a wide range of mixed Poisson distributions

Chapter 3

Maximum Likelihood Estimation For Finite Mixtures

3.1 Introduction

In the previous chapter we saw the derivation and the properties of several mixed Poisson distributions. We saw that, according to the choice of the mixing distribution, several mixed Poisson distributions can be derived. The estimation of the parameters of these mixed Poisson distributions is of special interest to researchers since such methods will allow the researcher to apply the described mixed distributions to real data.

In general the parameters of mixture models can be estimated in two distinct ways:

- The first one requires the parametric specification of the mixing distribution. Then the problem is just to estimate the parameters of the derived distribution. This approach is not always straightforward. The main reason is the intractability occurring due to the complicated nature of the mixed distribution. Consider a simple example: assuming a mixed Poisson model with a Gamma distribution as the mixing distribution we have to estimate the parameters of the resulting negative binomial distribution given in (2.22). This is relatively simple. But the case of the Poisson modified Bessel distribution given in (2.33) is clearly not so easy.

- The second is the nonparametric approach when we do not assume that the mixing distribution is of any specific form and we try to estimate the parameters of the mixture nonparametrically. Some authors (e.g. Lindsay and Roeder, 1995) use the term semiparametrically for these cases in order to show that in fact some kind of knowledge is incorporated in the estimation procedure. We adopt this terminology and we will call such models semiparametric models. This chapter is mainly devoted to such methods.

Finite mixtures play a very important role in mixture modelling. Assume that we know that the population consists of k subpopulations each one having a probability density function of some parametric form with different parameters say $f(x|\theta_j)$, j=1,...,k. Then the random variable X has probability density function (or probability function if X is a discrete random variable) of the form

$$f(x) = f_P(x) = \sum_{j=1}^{k} p_j f(x|\theta_j) \tag{3.1}$$

where $0 \le p_j \le 1$, j=1,...,k with $\sum_{j=1}^{k} p_j = 1$ and θ can be either a scalar (e.g. the Poisson distribution case) or a vector of parameters (e.g. the normal distribution case). The p_j's are called mixing proportions and they can be regarded as the probability that a randomly selected observation belongs to the j-th subpopulation. With P we denote the mixing distribution, which is the distribution that gives positive probability mass p_j at the points θ_j, j=1, . . . ,k and zero elsewhere. Assuming that $f(x|\theta)$ is the Poisson distribution, the vector of parameters θ reduces to the simple parameter ë of the Poisson distribution. We call the probability function given by

$$f_P(x) = \sum_{j=1}^{k} p_j \frac{\exp(-\lambda_j)\lambda_j^{\,x}}{x!} \tag{3.2}$$

with x= 0, 1, ..., and $\lambda_j > 0$, j= 1, . . ., k , as a k-finite mixed Poisson distribution with P as the mixing distribution. Note that finite Poisson mixtures are mixed Poisson distributions with a finite mixing distribution, and hence all the properties of the previous chapter apply. In order that finite Poisson mixtures are identifiable we need to impose the restriction $0 < \lambda_1 < \lambda_2 < \ldots < \lambda_k$ (Teicher, 1963). The reason is that otherwise interchanging the components would lead to the same mixture.

The importance of finite mixtures is dual. On the one hand a k-finite mixture is an appropriate model for describing populations which consist of k subpopulations. On the other hand, when we try to estimate the true probability function P with some probability function $\hat{P}$, we are confined to determine only a finite step distribution $\hat{P}$ as an estimate of P (Laird, 1978). Of course in this case we do not know a priori the number of support points of this distribution function. In both cases, we are restricted to estimate only a finite-step distribution. The number k of components of this finite mixture is either known a priori or it must be estimated from the dataset.

Both cases will be treated. Special care will be given to distinguishing between these two approaches, so as to avoid confusion. However, the two approaches (for known k and for unknown k) share some common elements which will be postulated. Note that the method for the case of unknown k contains several successive steps of the method for the case of known k.

Before describing the methods which will be used, it is interesting to mention three different sample types encountered in the analysis of mixture data. These sample types involve the presence or not of additional information for the subpopulation from which each observation comes.

At first we have to define the nature of our sample. According to Hosmer (1973a), it is possible to encounter three types of data. Following his presentation and notation we will refer to them as:

- M0 type : when we have observations only from the mixed distribution without any supplementary information. This is the most common case and the most difficult to handle. The majority of methods which we will describe assume such type of samples. Consider, for example, data referring to the number of accidents for the clients of an insurance company. If our sample simply contains the number of accidents with no other information, this sample is of type M0.

The next two sample types of samples contain observations from both the mixed distribution and from the component distributions. So, for some observations we have some knowledge about the subpopulation to which they belong (we refer to these data as known data). We can distinguish between the two types according to the information they contain about the mixing proportions. These types are :

- M1 type : when the sample contains both mixed and known data and when the known data contain no information about the mixing proportion. This is the case where we arbitrarily choose some members from each subpopulation without taking into account the mixing proportions. Such an example is the case where, from a file containing the number of accidents of persons from an insurance company we choose, say, n men and n women and a further number of other persons without sex identification. The arbitrary choice of n men and n women contains no information about the mixing proportion.

- M2 type : when the sample contains both mixed and known data and information about the mixing proportions is contained in the relative number of observations from the two components in the known data. For example, consider the case when we take randomly a

sample from the entire population and then we can recognise, for some observations, the subpopulations to which they belong. In this case we choose from the file of the insurance company a random sample of n persons and then we identify that m_1 of them are men and m_2 are women ($m_1+m_2=n$), and we also have other m cases without knowledge about the sex of the individuals. The ratio m_1 / n contains some information about the mixing proportion for the men.

In the sequel, we describe estimation procedures for the M0 sample. For the other two types the literature is very sparse. Hosmer (1973a) described the ML method for normal mixtures for a M2 sample while Murray and Titterington (1978) discussed minimum distance estimation from a M2 sample.

Intuitively, the analysis of samples of types M1 and M2 is easier since some additional information is available. For example, the widely used EM algorithm for mixtures needs to start from "good" initial values. Obviously very good initial values can be elicited from the known data.

This chapter is organised as follows. A detailed description of the ML method for finite mixtures is given. Although special emphasis is placed on finite Poisson mixtures, we try to describe the ML method for general mixtures. Some results for the ML method for mixtures are presented, which can lead to interesting improvements to the numerical methods used for deriving the estimates. They can also provide an insight into the method itself. We develop the ML estimation for Poisson mixtures arising from M2 type of samples, showing that the gain from the observations contain additional information can improve substantially the estimation procedure. The EM algorithm for mixtures is reviewed in depth in the sequel. This section contains a simulation experiment for selecting starting values which leads to clear and interesting results about the adequacy and the importance of the choice of initial values. We also propose an improvement of the algorithm which applies to all mixtures of the exponential family. The next part of this chapter is devoted to semiparametric ML estimation in mixture models. We review the proposed methods in a critical way and we try to demonstrate the advantages and disadvantages of such methods. Unfortunately our results discourage the use of such methods for Poisson mixtures for reasons which we try to explain from both the theoretical and the practical point of view.

3.2 The Maximum Likelihood Method for Finite Mixtures

The ML method is so far the most widely used method of estimation. For mixture models the ML method is again a very attractive approach. Actually, the associated efficiency and the well understood properties of the ML estimators are the reasons for the acceptance of the ML method as the most reliable method. For mixture models, the impact of computer intensive methods resulted in a very large number of applications based on ML estimation, especially since the later 60's. Previously, the computational difficulties in deriving the ML estimates had led the researchers to use different estimation methods.

In this section we provide a critical review of the ML method for mixture models. Special emphasis is given to finite mixtures of the Poisson distribution. We discuss the uniqueness of the ML estimators, algorithms for their calculation as well as their properties.

Interesting reviews of the ML method in mixture models can be found in Gupta and Huang (1980), Everitt and Hand (1981), Redner and Walker (1984), Titterington *et al.* (1985) and Bohning (1995).

3.2.1 The Likelihood Equations for Finite Mixtures

Suppose that we observe a random sample $X_1, X_2, \ldots, X_n$ (n is the sample size) where each X_i has probability function given by (3.1). Then the likelihood L of this sample is given by

$$L(P) = \prod_{i=1}^{n} f_P(x_i) = \prod_{i=1}^{n} \left(\sum_{j=1}^{k} p_j f(x_i | \theta_j) \right)$$

and the logarithm of the likelihood ℓ is given by

$$\ell(P) = \ln L(P) = \sum_{i=1}^{n} \ln \left(\sum_{j=1}^{k} p_j f(x_i | \theta_j) \right) \qquad (3.3)$$

Note that in (3.3) we express the loglikelihood as a function of the mixing distribution P, since in fact we want to maximise the likelihood over all the possible mixing distributions, i.e. over all the finite distributions with k-support points.

As usual, in order to find the ML estimators we need to equate all the partial derivatives of ℓ with 0, namely

$$\frac{\partial \ell}{\partial \theta_j} = \sum_{i=1}^{n} \frac{p_j}{f_P(x_i)} \frac{\partial f(x_i | \theta_j)}{\partial \theta_j} = 0 \qquad \text{for } j=1, \ldots, k \ , \qquad (3.4)$$

$$\frac{\partial \ell}{\partial p_j} = \sum_{i=1}^{n} \frac{f(x_i|\theta_j) - f(x_i|\theta_k)}{f_P(x_i)} = 0 \qquad \text{for } j=1, \ldots, k-1 \qquad (3.5)$$

(3.4) and (3.5) must be solved together to obtain the ML estimates. Clearly, this a very difficult task. Equations (3.4) and (3.5) are the likelihood equations for finite mixture of a general probability density function (or probability function) $f(x|\theta)$. For specific choices of $f(x|\theta)$ the likelihood equations may be simplified. For simplicity we treat θ as a scalar.

Finding a simple analytical solution for these equations is hopeless and the difficulty in solving all of these equations was the main reason why the ML method was not used in mixture models till the end of the 60's. The need of numerical methods is obvious. A first choice is the well known algorithms for solving a non-linear system of equations such as the Newton Raphson or methods designed to minimise certain functions such as the one based on the Fletcher-Reeves algorithm. However, the larger the number of parameters to be estimated the more difficult the applicability of these methods. The main problems with some of them are the failure to converge and that they can be very slow because at each iteration a large matrix has to be calculated and inverted. Everitt (1984b) and Atwood *et al.* (1992) compared some of them in a normal mixture problem. See also the papers of Dick and Bowden (1973), Hosmer (1973b) and Peters and Walker (1978), among others.

Razzaghi and Rayens (1987) developed a modified ML method for estimating the mixing proportions of a mixture with known components. This estimator is given in closed form. Kazakos (1977) gave an iterative algorithm for deriving the ML estimates for the mixing proportions.

Apart from the known numerical methods for solving a system of non-linear equations an iterative scheme is also available for finding the ML estimates in the mixture case. Such a method was proposed by Hasselblad (1966) for grouped data from normal mixtures and by Hasselblad (1969) for mixtures of members of the exponential family. Independently, Behboodian (1970) described the same algorithm for normal mixtures. This iterative algorithm was in fact an EM algorithm, formally discussed in Dempster *et al.* (1977). It is interesting to note that the original derivation of the algorithm by Hasselblad (1969) and Behboodian (1970) was based on the likelihood equations and not on the notion of 'missing data' which is the key ingredient of the EM algorithm. We will describe the EM algorithm for mixtures in a subsequent section.

Some problems may occur in applying the ML method in finite mixtures. These are the following:

- The likelihood in mixture models may be unbounded from above and hence it is not possible to obtain the ML estimator. This occurs in the case of normal mixtures when the variance of each component is unknown. Then, when the method tries to fit a component with only one observation the variance of this component cannot be calculated (Hathaway, 1985). In this case a constrained likelihood ought to be used, imposing constraints which do not allow the variances to be less than a small value (Hathaway 1986a).

- Basford and McLachlan (1985) demonstrated some cases of normal mixtures where the ML estimates must be handled with care, because of the existence of several local maxima in several different neighbourhoods of the parameter space. Hawkins (1972) treated the case of multiple maxima in normal mixtures. The multiplicity of maxima results from the inappropriateness of the model. As seen in the sequel, in real applications it is common that the likelihood can not increase any further by adding a new component. So, if we try to add a redundant component, the estimation will result in multiple maxima, which correspond to components with a zero mixing proportion (or with a proportion very close to zero due to numerical perturbations) or components which are close together.

- EM algorithm can be very slow in cases of multivariate finite mixtures. To improve its performance some alternations have been proposed (see, e.g., Pilla and Lindsay, 1996 and Oskrochi and Davies, 1997).

For Poisson mixtures the likelihood function is bounded and the maximum exists.

Lindsay (1983a) gave conditions which have to be satisfied from the ML estimates. In order to describe these conditions we need to introduce some notation.

Let $D(G,P)$ be the directional derivative of the loglikelihood from the mixing distribution P at the direction of another mixing distribution G.

$$D(G,P) = \lim_{e \to 0} \left[\frac{\ell\big((1-e)P + eG\big) - \ell(P)\big)}{e} \right].$$

This quantity measures the infinitesimal change of the loglikelihood when a new distribution G is added to the mixing distribution P. Of special interest is the case when the new mixing distribution G is a degenerate distribution at the point è. Then we can define the gradient function

$$D(\theta, P) = \sum_{i=1}^{n} \left\{ \frac{f(x_i|\theta)}{f_P(x_i)} - 1 \right\} \quad . \qquad (3.6)$$

$D(\theta, P)$ plays an important role in the case of semiparametric ML method for mixtures. Using the gradient function we can state the following theorem of Lindsay (1983a) which provides sufficient conditions for an estimator $\hat{P}$ to be the ML estimator of the mixing distribution P in the case of a known number of support points.

Theorem 3.1 (Lindsay, 1983a). $\hat{P}$ is the restricted ML estimator iff, for each support point θ^* of $\hat{P}$, the following relations hold:

a) $D(\theta^*, \hat{P}) = 0$

b) $\left.\frac{\partial D(\theta, \hat{P})}{\partial(\theta)}\right|_{\theta=\theta^*} = 0$ and

c) $\left.\frac{\partial^2 D(\theta, \hat{P})}{\partial(\theta)^2}\right|_{\theta=\theta^*} - \sum_{i=1}^{n}\left(\frac{f(x_i|\theta^*)}{f_P(x_i)}\right)^2 \leq 0$

given that $f(x|\theta)$ is twice differentiable.

The proof of the above theorem is based on the likelihood equations.

The importance of this theorem is revealed in section 3.7 when a similar theorem for the semiparametric case will be given.

3.2.2 A Result for the Maximum Likelihood Estimation for Finite Mixtures from the Exponential Family

In this section we provide a proof of an interesting result for general finite mixtures from the exponential family. An alternative proof was given earlier by Lindsay (1981) (which seems to have passed unnoticed. The derivation is given in a general form to include all the members of this family, continuous or discrete.

In order to obtain the ML estimators for the k-finite mixture model, we have to solve the system of (3.4) and (3.5). We show that if $f(x|\theta)$ belongs to the one-parameter exponential family the ML estimators satisfy the first moment equation. Some of the best known distributions belong to this family and hence the results can be applied to many cases where finite normal mixtures, finite exponential mixtures and finite Poisson mixtures among other models are appropriate.

Let us suppose that the density $f(x|\theta)$ comes from the one-parameter exponential family, namely that $f(x|\theta)$ can be written in the form:

$$f(x|\theta) = \exp\left[\theta x c + h(x) - k(\theta)\right] \quad , \tag{3.7}$$

where c is some constant and the functions $h(x)$ and $k(\theta)$ depend only on x and θ, respectively. It can easily be shown that $E(X) = \mu(\theta) = k'(\theta)/c$ and that $Var(X) = \sigma^2(\theta) = k''(\theta)/c^2$. Furthermore, the first derivative with respect to è is of the form:

$$f'(x|\theta) = f(x|\theta)\left(cx - k'(\theta)\right) = cf(x|\theta)\left(x - \mu(\theta)\right) \quad . \tag{3.8}$$

We now prove the following theorem.

Theorem 3.2 For finite mixtures from the one parameter exponential family defined in (3.7), one of the ML equations is the same as the first moment equation, and hence the ML estimates ought to satisfy the first moment equation.

Proof:

The estimating equations for the general finite mixture model are given in (3.4) and (3.5). Multiplying the i-th equation in (3.4) by p_j , $j=1, 2, \ldots, k$, and adding the resulting equations we obtain

$$\sum_{i=1}^{n} \frac{\left[f_P(x_i) - f(x_i|\theta_k)\right]}{f_P(x_i)} = 0 \quad \text{or, equivalently}$$

$$\sum_{i=1}^{n} \frac{f(x_i|\theta_k)}{f_P(x_i)} = n \quad . \tag{3.9}$$

On the other hand, it follows from (3.5) that

$$\sum_{i=1}^{n} \frac{f(x_i|\theta_j)}{f_P(x_i)} = \sum_{i=1}^{n} \frac{f(x_i|\theta_k)}{f_P(x_i)} \quad , \qquad j=1, \ldots, k \quad . \tag{3.10}$$

From (3.9) and (3.10) it may be concluded that the ML estimates satisfy the relation

$$\sum_{i=1}^{n} \frac{f(x|\theta_j)}{f_P(x_i)} = n \quad , \qquad j=1, \ldots, k \quad . \tag{3.11}$$

Note that equation (3.11) is equivalent to condition (b) of Theorem 3.1, while condition (a) of Theorem 3.1 is equivalent to equation (3.4).

Substituting in (3.5) for $f'(x_i|\theta_j)$ as given by (3.8) we obtain

$$\sum_{i=1}^{n} \frac{f(x_i|\theta_j)}{f_P(x_i)}\left(x_i - \mu(\theta_j)\right) = 0 \quad , \qquad j = 1, \ldots, k \quad . \tag{3.12}$$

Then by setting

$$w_{ij} = \frac{f(x_i|\theta_j)}{f_P(x_i)} \qquad i = 1, \ldots, n, \;\; j = 1, \ldots, k \qquad (3.13)$$

equation (3.12) can be written as

$$\sum_{i=1}^{n} w_{ij}\left(x_i - \mu(\theta_j)\right) = 0\,, \qquad j=1, \ldots, k \qquad . \qquad (3.14)$$

If we consider mean value reparametrization for the density $f(x|\grave{e})$ and solve for the mean value parameters we obtain from (3.14) that

$$\mu(\theta_j) = \frac{\sum_{i=1}^{n} w_{ij} x_i}{\sum_{i=1}^{n} w_{ij}} \qquad , \qquad \text{j=1,. . ., k .}$$

This, using (3.10) is equivalent to

$$\mu(\theta_j) = \frac{\sum_{i=1}^{n} w_{ij} x_i}{n} \qquad , \qquad \text{j=1, . . ., k} \qquad . \qquad (3.15)$$

The latter implies that the ML estimates of the mean value parameters can be written as weighted sample means.

Suppose now that the ML estimators for the parameters $\mu(\theta_j)$, $j=1, \ldots, k$ have been calculated from (3.15). It is well known that the mean of a mixture is the weighted mean of the means of all components weighted by the mixing proportions. Then, from (3.1) we estimate the mean $E(X)$ of the finite mixture as

$$E(X) = \sum_{j=1}^{k} p_j \mu(\theta_j) = \sum_{j=1}^{k} \frac{\sum_{i=1}^{n} w_{ij} x_i}{n} p_j = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n} w_{ij} x_i p_j}{n} =$$

$$= \frac{\sum_{i=1}^{n} \frac{\sum_{j=1}^{k} p_j f(x_i|\theta_j)}{f_P(x_i)} x_i}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x},$$

where $\overline{x}$ is, as usual, the sample mean.

Hence, the ML estimates of the mean value parameter of a k-finite mixture from the one-parameter exponential family coincide with the first moment equation. So, we have another family of distributions which satisfy the first moment equation. This is also true for

members of the power series family of distributions (see, e.g., Johnson *et al.*, 1992). Sprott (1983) showed that this result holds for the convolution of two power series distributions as well as for compound (or generalised) distributions of members of the power series family. A generalisation of the power series family shares the same property, as Kemp (1986) showed. It is interesting that some of the most well known distributions belong to the one-parameter exponential family, like the Poisson, the normal, the exponential, the Gamma and other distributions. For many of them the parameter è represents the mean of the distribution. Behboodian (1970) has shown a similar result for finite normal mixtures.

Note that equation (3.14) was the basis for the iterative algorithm of Hasselblad (1969) and Behboodian (1970).

The above result can find interesting applications for improving the EM algorithm for finite mixtures as it will be illustrated in the sequel.

3.2.3 The Variance Covariance Matrix for the Case of 2-Finite Mixtures

It is well known that the variance-covariance matrix for the ML estimates is the inverse of the information matrix, **I** whose ij-th element is given by:

$$I_{ij} = -nE\left(\frac{\partial^2 \ln(f_P)}{\partial\theta_i \partial\theta_j}\right) = -nE\left(H_{ij}\right)$$

where

$$H_{ij} = \frac{\partial^2 \ln(f_P)}{\partial\theta_i \partial\theta_j}.$$

We now give the elements of the matrix **H** for the case of k-finite Poisson mixtures defined by (3.2).

Matrix **H** is a (2k-1)×(2k-1) matrix which can be rewritten as

$$\mathrm{H} = \begin{bmatrix} \mathrm{S} & | & \mathrm{R} \\ -- & | & -- \\ \mathrm{R}^T & | & \mathrm{T} \end{bmatrix}$$

where $\mathbf{R}^{\mathrm{T}}$ denotes the transpose of **R.** The submatrices **S, R** and **T** are calculated as follows.

Matrix **S** is a $(k-1)\times(k-1)$ matrix with elements

$$S_{ij} = \frac{\partial^2 \ln(f_P(x))}{\partial p_i \partial p_j},$$

i.e. **S** is the matrix whose elements are the covariance terms from the mixing proportions.

Matrix **T** is a $k \times k$ matrix with elements

$$T_{ij} = \frac{\partial^2 \ln(f_P(x))}{\partial\theta_i \partial\theta_j} ,$$

i.e. **T** is the matrix whose elements are the covariance terms from the parameters of the component distribution. Finally, matrix **R** is a $(k-1)\times k$ matrix with elements

$$R_{ij} = \frac{\partial^2 \ln(f_P(x))}{\partial p_i \partial\theta_j} .$$

The elements of each matrix are given for the case of a k-finite Poisson distribution of (3.2) by the following formulas:

$$T_{ij} = \frac{-p_i p_j \left[f_j(x-1) - f_j(x)\right]\left[f_i(x-1) - f_i(x)\right]}{\left(f_p(x)\right)^2} \qquad \text{for } i, j = 1, \ldots, k \text{ and } i \neq j$$

$$T_{jj} = \frac{p_j\left[f_j(x-2) - 2f_j(x-1) + f_j(x)\right] - p_j^2\left[f_j(x-1) - f_j(x)\right]^2}{\left(f_P(x)\right)^2}$$

for the diagonal elements

$$S_{ij} = -\frac{\left(f_j(x) - f_k(x)\right)\left(f_i(x) - f_k(x)\right)}{\left(f_P(x)\right)^2} \qquad \text{for } i, j = 1, \ldots, k\text{-}1$$

and

$$R_{ij} = -\frac{p_j\left(f_j(x-1) - f_j(x)\right)\left(f_i(x) - f_k(x)\right)}{f_P^2} \quad \text{for } i = 1, \ldots, k\text{-}1,\ j = 1, \ldots, k \text{ and } i \neq j$$

and elements

$$R_{jj} = \frac{\left(f_j(x-1) - f_j(x)\right)g(x) - p_j\left(f_j(x-1) - f_j(x)\right)\left(f_j(x) - f_k(x)\right)}{\left(f_P(x)\right)^2},$$

for $j = 1, \ldots, k\text{-}1$

where $f_i(x) = f(x|\theta_i) = \exp(-\theta_i)\theta_i^{\,x} / x!$, i=1, . . , k.

The elements of the information matrix are calculated as

$$I_{ij} = -n\sum_{i=1}^{n} H_{ij} f_P(x)$$

The inverse of the matrix **I** is the variance-covariance matrix for the ML estimates.

The ML estimates for finite mixtures suffer from large variances especially when the components are not well separated (see, e.g., Hasselblad , 1969). Basford *et al.* (1997)

compared the asymptotic results with bootstrap standard errors for a normal mixture model, showing that bootstrap standard errors are more accurate.

3.3 The EM Algorithm for Finite Mixtures

3.3.1 The EM Algorithm

The EM algorithm is a powerful tool for ML estimation. Its widespread use is phenomenal. Since its first appearance in Dempster *et al.* (1977) it has been used in a variety of contexts and applications. EM stand for the Expectation and Maximisation steps, which are the two basic steps of iterative algorithms.

Roughly speaking, the key ingredient of the EM algorithm is the 'missing data' principle. To put it another way, we treat our dataset as having 'missing values' even in the case when there are not really missing values. If these 'missing values' had been observed, the estimation would be simple. So, the EM proceeds by estimating these 'missing values' by their expectation conditionally on the current estimates (the E-step) and then it uses these expectations to maximise the complete likelihood (M-step).

More formally, suppose that our full data representation Y contains an observed part and a missing part (or a part which can be considered as missing), i.e. $Y_i = (X_i, Z_i)$, where X_i is the observable part of our data and the Z_i is the missing part. Then the EM algorithm iterates between the following two steps:

E-step: Calculate the expected values for the Z_i given the data and the current estimates of the parameters of interest and

M-step: Maximise the complete likelihood using the expected values of the missing data calculated at the E-step and the data, to find the new estimates.

The convergence of the EM algorithm has been treated by Wu (1983) and Meilijson (1989). Unfortunately, convergence at the global maximum cannot be ensured, and often local maxima as well as saddle points can be found.

Despite these problems the EM algorithm is a very useful algorithm for calculating the ML estimates in a wide variety of applications, including missing data problems or problems which can be considered as missing data. More details for the EM algorithm can be found in the original paper of Dempster *et al.* (1977). Recent improvements of the EM algorithm are presented in Meng and Van Dyk (1997) and the discussion therein. We focus our attention on the EM algorithm for finite mixture models.

3.3.2 The EM Algorithm for Finite Mixtures

Standard numerical techniques were needed for calculating the ML estimates, till the introduction of the EM algorithm which facilitates the required calculations. Hasselblad (1969) suggested a method for obtaining ML estimates for the case of k-finite mixtures of distributions which belong to the exponential family. This family includes many of the most popular distributions. Independently, Behboodian (1970) came to the same result for only the case of the normal distribution. Later, these algorithms were identified as EM type algorithms.

The missing data representation is suitable for mixture models since we may consider that the complete data specification $\mathrm{Y_i}$ is represented as $Y_i = (X_i, Z_i)$ where $\mathbf{Z}_i$ is a 1 x k vector of elements

$$Z_{ij} = \begin{cases} 1, & \text{if the observation } \mathrm{X_i} \text{ belongs to the j-th subpopulation} \\ 0, & \text{otherwise} \end{cases} .$$

If we knew the real values of these vectors the maximisation would be simple. With such a representation the complete loglikelihood is of the form:

$$\ell_c = \sum_{i=1}^{n} \sum_{j=1}^{k} Z_{ij} \ln(p_j) \ln\left(f(x_i|\theta_j)\right).$$

Thus, following the general EM formulation, at the E-step we have to estimate the 'missing data' Z_{ij} and at the M-step we have to maximise the complete loglikelihood ℓ_c which is rather simple as it will be seen.

We now concentrate on the Poisson case.

The method starts with initial values and calculates at every step new estimates for both p_i and, $\ddot{e}_i$, $i = 1,2, \ldots, k$.

At each step, the estimates can be calculated from the following simple scheme:

Step 1 (E-step) Given the current values for λ_j^{old} , $j = 1, \ldots, k$ and p_j^{old} , $j=1, \ldots, k$ calculate the probability w_{ij} that the observation X_i belongs to the j subpopulation after observing it, i.e. the posterior probability of X_i belonging in the j subpopulation.

$$w_{ij} = \frac{p_j^{old} f(x_i|\lambda_j^{old})}{f_P(x_i)} \tag{3.16}$$

Step 2 (M-step) Calculate the new estimates as

$$\lambda_j^{new} = \frac{\sum_{i=1}^{n} w_{ij} x_i}{\sum_{i=1}^{n} w_{ij}} \qquad \text{for } j=1, 2, \ldots, k \qquad (3.17)$$

and

$$p_j^{new} = \frac{\sum_{i=1}^{n} w_{ij}}{n} \qquad \text{for } j=1, 2, , \ldots, k \quad . \qquad (3.18)$$

Step 3 Check if some condition is satisfied in order to terminate the iterations, otherwise go back to step 1, putting the currently estimated values for p's and ë's as the initial values.

The above scheme is used for the case of finite Poisson mixtures. The general scheme for mixtures from the exponential family covering the normal, the Gamma, the Poisson and the binomial distributions can be found in Hasselblad (1969), while a more general scheme can be found in McLachlan and Basford (1988).

We caution that the above described scheme is not the only one which has appeared in the literature. The original derivation of Hasselblad (1969) is slightly different, mainly with respect to the representation of the two steps. For example we can use (3.13) to define the weights. Equation (3.16) differs from (3.13) only with respect to the term p_j appearing in the nominator of (3.16). This quantity vanishes in (3.17) and (3.18) as appearing in both the numerator and the denominator. We preferred this description mainly because it relates to the general EM formulation. Note also that for some applications, like cluster analysis, the weights w_{ij} (i.e. the posterior probabilities that the i-th observation belongs to the j-th subpopulation) themselves are of interest (see, e.g., Symons *et al.*, 1983).

Behind the missing data derivation of the EM algorithm, the algorithm can be described as follows:

At step 1 we just obtain the weights for each observation using the current observations. These weights are used to separate the entire sample in k subsamples in the sense that each observation X_i belongs to the j-th subsample as $Y_{ij} = X_i w_{ij}$. At step 2 we estimate the parameters of the distribution (Poisson in our case) for each sample $\mathrm{Y}_j = \left(Y_{1j}, Y_{2j}, \ldots, Y_{nj}\right)$. For distributions with ML estimators which can be written in closed form, step 2 is easy to be carried out. Otherwise step 2 can be carried out numerically.

If we use matrix notation we can define at step 2 the $n \times k$ matrix **W** with its ij-th element equal to w_{ij} . Suppose also that our observations can be represented by the vector $\mathbf{X^t}$ = (X_1 ,X_2 ,...,X_n). If $\mathbf{1}_n$ denotes the vector with all of its n elements equal to 1, then the vectors $\boldsymbol{\theta} = (\lambda_1, \lambda_2, \ldots, \lambda_k)$ and $\mathbf{p} = (p_1, p_2, \ldots, p_k)$ of the parameters can be obtained as

$$\boldsymbol{\theta} = \mathbf{X}^t \mathbf{W}\mathbf{A}^{-1} \tag{3.19}$$

where **A** is the diagonal matrix with diagonal elements the elements of the vector $\mathbf{1}^t\mathbf{W}$, and

$$\mathbf{p} = n^{-1}\mathbf{A} \tag{3.20}.$$

The usefulness of this representation is that it can be easily extended to both the bivariate case and to the case of M2 sample type, as described in the sequel.

If k=1 the solution is the known result that the mean of the sample is the ML estimator for the parameter ë of the Poisson distribution. The estimates for p_i's in each step are simply the means of the posterior probabilities.

Hathaway (1986b) showed that the EM algorithm for mixtures can be interpreted as a method of co-ordinate descent on a particular objective function.

The general description of the algorithm as an EM algorithm provides evidence about the convergence of the algorithm.

The use of the EM algorithm for ML estimation in finite mixtures has some disadvantages:

- Different initial values lead to different estimates and we cannot be sure as to whether we obtained the global maximum or a local maximum. So, we have to check it and try with a variety of initial values. The conditions which must be satisfied by the global maximum are given in Theorem 3.1. According to Wolfe (1971) the existence of multiple maxima may be attributed to the fact that several k-component separations can be obtained. For example a population can be divided into subpopulations by sex, by age, by social status etc. So, the solution we obtain depends on how close to each of these ways of dividing the population the initial values are.

- The method is very slow, requiring a large number of iterations to achieve some kind of convergence.

- The variances of the ML estimators may be very large if the different parameters for the distinct components are too close. In other words the algorithm has a difficulty in recognising the different components when they are not much separated.

•Another drawback is when we need a good "stopping rule" in order to stop the iterations. The method is very sensitive in the sense that different "stopping rules" can lead to quite different estimates. This is caused by the fact that at every iteration the loglikelihood increases by a very small amount and at the same time the estimate changes drastically.

A lot of attempts to cope with such problems have been made and in the sequel we discuss these issues. However, we have to adhere to this method its simplicity and its efficiency if handled with care. Another advantage of the method is that the ML estimate lies in the admissible range whenever the initial guess is in the same range and this is not the case with other methods, including numerical methods for solving the likelihood equations. The algorithm is easily programmable on any computer, using commonly used statistical packages.

The asymptotic variances of the estimators can be calculated in the usual way by constructing and inverting the information matrix. Louis (1982) extracted the observed information using the EM algorithm for a normal mixture.

Another interesting point for the EM algorithm is that it may be used in conjunction with some other methods. For example, the slow convergence of the EM algorithm is a problem. The Newton iterative scheme is much quicker in convergence but it needs very precise initial values. So, a plausible solution might be : Make some iterations with the EM algorithm to nearer the maximum and then locate it using the Newton method (or any other method). The idea is also described in Aitkin and Aitkin (1996).

In the sequel we consider an application of the EM algorithm to real data.

Example 3.1 The data in Table 3.1 represent the number of crimes committed in a one month period in Greece from January 1982 until January 1994 (145 observations). The dataset shows a large amount of overdispersion ($\bar{x}$ = 2.2413, s^2 = 3.3833). It is therefore reasonable to assume that they come from a mixed Poisson distribution. The least inhomogeneity can be incorporated in the model by assuming a 2-finite mixture model. The ML estimates for the parameters were estimated via the EM algorithm.

Table 3.1
The number of crimes committed in one month in Greece for the period January 1982 to January 1994

x	0	1	2	3	4	5	6	7	8	9
observed frequency	21	41	32	16	19	9	8	1	2	1

Source: The Greek newspaper 'TA NEA' 15/2/1994

The estimated parameters with their standard errors in parentheses, were p_1=0.672 (0.047), $ë_1$=1.488 (0.143), $ë_2$=3.788 (0.312). The standard errors were calculated via jacknifing. The reason is to enable comparisons with other methods which will be applied in the sequel. More details about jacknife estimates of standard errors can be found in the Appendix .

3.3.3 The Choice of Initial Values for the EM Algorithm

3.3.3.1 A review

The initial values are of great interest in the implementation of the EM algorithm (and not only for the EM). Laird (1978) proposed grid search for setting the initial values. In practice it is useful only for large numbers of support points. Otherwise, the time required for the grid search can be spent for starting the EM algorithm from several different sets of initial values. Leroux (1992) suggested to use supplementary information in order to form clusters and then to use the mean of each cluster as initial values. For example, if the data refer to accidents and supplementary information is available for each individual , e.g., age or sex, we can form clusters of individuals with the same characteristics from which to obtain the initial estimates.

Finch *et al.* (1989) proposed that for a 2-finite normal mixture only the mixture proportion ought to be given an initial value and thus the rest of the parameters to be calculated automatically by this value. The idea is that, given the mixing proportion p, we separate the sample so that the first [np] observations belong to the first component and hence, we take the mean of these observations as the initial value for the mean of the first component while the remaining observation are considered to belong to the second component taking their mean as the initial value for the mean of the second component ([a] stands for the integer part of a).

Woodward *et al.* (1984) proposed a clustering method for obtaining initial values. Given the mixing proportion and separating the sample in successive subsamples the initial values are taken by minimising the within clusters sum of squares.

Bohning *et al.* (1994) proposed to start with well separated values because their experience showed that with such initial values the algorithm can converge easier.

Another natural choice is to begin with estimates from other much simpler methods, like the moment method. Lindsay and Basak (1993) advocated such starting values for multivariate normal mixtures, ignoring however that it is highly possible that moment estimates do not exist in the sense that the moment equations lead to estimates outside the admissible range (e.g. negative Poisson parameters) This issue is examined in depth in chapter 4. Furman and Lindsay (1994a, 1994b) suggested the use of moment estimates as initial values, carrying out a small simulation experiment showing that the moment estimates are usually near the maximum and thus they are reliable either as initial values or as close approximations of the ML estimates. Our results of section 3.2.2 explain this issue.

Fowlkes (1979) proposed some graphical and ad-hoc methods for choosing initial values in the case of normal mixtures. For a more thorough treatment of graphical methods one can refer to the book of Titterington *et al.* (1985).

McLachlan (1988) presented a method for selecting initial values for the case of multivariate mixtures. The idea is to use principal components analysis in order to obtain a naive clustering of the data and then choose the initial values on the basis of the clusters considered.

3.3.3.2 A Simulation Comparison

In the sequel, a small simulation comparison of some methods for setting initial values is given for the case of 2-finite Poisson mixtures. For k>2 the choice of initial values is more difficult and will not be treated. Note that many of our methods will not work properly for k>2, or at least they will be difficult to handle. However, the results from 2-finite mixtures can give useful insight for these cases too. In this simulation experiment we used the following methods for choosing initial values:

a) the 'true' values (TR) of the parameters. Of course in practice we are not able to know these values but since for several simulation studies the true values are known we include these initial values to this comparison. Note that if we want simply to analyse a data set, the impact of computer resources would be enhanced by using several initial values with very little effort and usually with little computing time. The problem is crucial in simulation studies where a large number of repetitions is needed. Consider, for example, the bootstrap approach of the Likelihood Ratio Test (McLachlan, 1987) which is examined in depth in chapter 6. A great number of samples ought to be taken for constructing the distribution of the

test statistic for a specific k-finite Poisson mixture. In this case the true values are known. This leads to considering the true values as candidates in choosing the initial values. It is admitted that the results derived are optimistic about the actual operating characteristics. Moreover there are several theoretical concerns about what 'true' really means. We include these starting values in our simulation in order to compare their computational characteristics with the rest methods

b) moment estimates (MOM). A complete description of the calculation of these moment estimates can be found in the section for the moment estimation method, given in chapter 4.

c) The method is slightly different from the method proposed by Finch *et al.* (1989). Instead of setting arbitrarily the mixing proportion we calculate it from the dataset. By this method we find the mean of our data, and then we consider the initial value of the parameter p_1 to be the proportion of observations lower than the mean. The mean of all these observations is the initial value for $ë_1$ while the mean of the rest of the observations is the initial value for $ë_2$. If the initial value for $ë_1$ is 0 we set $ë_1$=0.1. This method is referred to as (F).

d) This method finds the initial value for p_1 as in the previous method. Then the initial values for $ë_1$ and $ë_2$ are calculated as follows:

$$\lambda_1 = \bar{x} - \sqrt{\frac{\left(s^2 - \bar{x}\right)(1-p_1)}{p_1}} \quad \text{and} \qquad \lambda_2 = \bar{x} + \sqrt{\frac{\left(s^2 - \bar{x}\right)p_1}{(1-p_1)}}$$

where $\bar{x}$ is the sample mean and s^2 is the sample variance. The motivation for this algorithm is the fact that we match with the initial values the mean and the variance of the sample, i.e. the initial estimates satisfy the first two moment equations. If $ë_1 \le 0$ we set $ë_1=0.1$. This method is referred to as (MF).

5) We set p_1=0.5 and $\lambda_1 = \bar{x} - s$, $\lambda_2 = \bar{x} + s$. This initial guess is symmetric, and also satisfies the first two moment equations of the observed dataset. The choice of the value 0.5 as an initial value for the mixing proportion is expected to work reasonably only when the mixing proportion is near 0.5. This method however can be easily extended to more than two components. The method is referred to as (HY).

Other methods were also included in the simulations, but since they were inferior to those discussed above we will not report their results. Such methods were:

•A method which separates the interval $[0, d]$ in three subsets of equal size, where $d = \max(X_i)$, and then chooses as initial values $\lambda_1 = \frac{d}{3}$, $\lambda_1 = \frac{2d}{3}$ and $p_1 = 0.5$. This method is discussed in Bohning *et al.* (1994).

•The grid search of Laird (1978) and others.

In order to examine the performance of the above defined methods for setting initial values, we need some criteria. We used the following criteria:

1. At first, we need to know if the initial values assumed exist. For example, the moment estimates do not exist very often. In fact, we want a method which is able to provide us with initial estimates in as many cases as possible. From the above methods only the TR and F methods always give initial estimates. The remaining methods can fail to provide us with initial estimates.

2. Another criterion is the speed of convergence which is measured by the number of iterations until convergence is attained. The cost for building up the initial guess is usually negligible. Usually it requires less than the computing time for one iteration of the EM algorithm and hence it will not be taken into account. The mean number of iterations until convergence will be reported using the same convergence criterion for all the methods.

3. Another necessary ingredient for a good initial guess is the ability of this guess to lead to the global maximum. The likelihood surface for mixture models is known to have many local maxima which are not global. A general strategy to avoid obtaining a local maximum is to start from several initial values. We examined whether the initial values considered were able to locate the global maximum. To check if the global maximum has been obtained we used the following method: From all the methods applied to the given sample, we obtained the different values of the maximised loglikelihoods. Then we obtained the maximum value of all the methods, say the value $L_{\max}$. If the value of the maximised loglikelihood for method i was 0.1% far from $L_{\max}$, then we reported this case as failure to obtain the global maximum. Such an approach admits that the global maximum has been obtained by at least one of the methods. The effect of this is that the probability that the global maximum has been obtained is overestimated. However, this is negligible, and it does not cause serious problems in this comparative study.

From the above discussion we can deduce that an excellent method for obtaining the initial values can be described as a method which never fails to give initial values and always

converges to the global maximum after very few iterations. Such a method does not exist in general.

These three criteria were used for assessing the performance of the methods. Tables 3.2-3.10 contain the results from the simulation experiment. In this experiment we simulated 1000 samples from several distributions, and all the methods available (TR, MOM, F, MF, HY) were used to obtain initial estimates. The mean number of iterations until convergence, the proportions of times the method failed to provide initial values and the proportions of times the method converged to the maximum are reported.

The distributions used were 2-finite Poisson mixtures, negative binomial and 3-finite Poisson mixtures. For the first case the true values were used as initial estimates while for the other two this was not possible.

Looking at Tables 3.2-3.4 we can see that the MOM method is usually the method which requires less iterations. Only the TR method can compete on this issue in the case where we sampled from the 2-finite Poisson mixture. Unfortunately, the MOM method has a high probability of failing to provide initial estimates, especially with small sample sizes and not well separated distributions as seen in Tables 3.5-3.7. The TR method works very satisfactorily when the components are well separated. As far as the rest of the methods are concerned the F method seems to be preferable because it never fails to provide initial estimates and requires fewer iterations in comparison with the MF and HY methods. The HY method is reasonable only when the true mixing proportion is near 0.5.

A very interesting finding is that the TR method never failed in locating the global maximum when the data were generated from a 2-finite Poisson mixture. This provides a good strategy for simulation studies : To use these values only. Clearly we are almost certain that the global maximum will be obtained and thus no other initial values are needed. In fact the probability of failure will be negligible as our simulations reveal.

Let us now examine the case where the data do not come from a 2-finite Poisson mixture but we try to estimate the parameters of a 2-finite Poisson mixture. In this case the MOM method is again very attractive because of the smaller number of iterations usually needed. However the F method is an interesting competitor in this case. For distributions with high overdispersion it requires a few iterations and it possesses the appealing property that exists in all cases. However due to the high proportion of times that all the methods failed to obtain the global maximum, a combination of them would be appropriate in order to increase the probability of obtaining the global maximum.

The HY and MF methods were inferior to the MOM and F methods in all the cases, and they also have a great chance to fail in providing us with initial values. As far as the MF method is concerned this failure was not expected. This method is a modified version of the F method, starting from points satisfying the two first moments of the dataset, while the F method is a heuristic one. The HY method has only one advantage: it is easily extended to cases with more components, and hence its performance is interesting. All the other methods cannot be easily extended to cases with more components, with the exception of the MOM method. However, the MOM method is not trustworthy for more components because it usually fails to give initial values (see section 4.3).

In concluding we may say that if the true values are known (e.g. in simulation studies) then it is better to use them as initial values.

In practical situations when the true values are not known, the MOM method is preferable, especially if the data show substantial overdispersion and the sample size is not small. Otherwise the F method is suggested. It is interesting that all the methods show an ability of obtaining the global maximum. The large proportion of the failure to obtain the global maximum for the case of sampling from a 2-finite Poisson mixture first results from the underdispersion of the data (which is probable when sampling from a distribution with very small overdispersion). As will be seen in Chapter 6, when the sample variance is smaller than the sample mean, the simple Poisson distribution is the appropriate distribution and, thus, there are multiple global maxima when trying to find the ML estimates for a 2-finite Poisson mixture.

Table 3.2
The mean number of iterations for all the methods when sampling from a 2-finite Poisson mixture

	p_1=0.5, λ_1=1, λ_2=2 sample size				p_1=0.5, λ_1=1, λ_2=8 sample size			
method	50	100	250	500	50	100	250	500
TR	114.90	157.54	188.55	180.95	7.65	7.02	6.45	5.91
MOM	84.86	116.41	139.11	125.24	8.37	7.74	7.17	6.64
MF	133.34	164.20	181.59	176.60	8.51	8.42	8.59	8.47
HY	141.94	171.96	186.80	178.26	8.76	8.04	7.42	6.85
F	105.90	150.53	179.03	174.04	8.20	8.16	8.33	8.24
	p_1=0.8, λ_1=1, λ_2=5 sample size				p_1=0.1, λ_1=1, λ_2=8 sample size			
	50	100	250	500	50	100	250	500
TR	25.03	19.32	15.11	13.52	14.53	10.73	9.19	8.25
MOM	29.73	26.75	24.49	23.88	23.48	17.53	12.39	11.30
MF	31.35	29.07	28.22	28.15	44.35	36.37	30.40	29.49
HY	41.28	37.50	34.33	33.21	44.50	36.44	30.42	29.31
F	31.09	29.46	28.22	28.08	41.62	34.28	28.95	28.26

Table 3.3
The mean number of iterations for all the methods when sampling from a negative binomial distribution

	n=1, p=0.5 sample size				n=2, p=0.25 sample size			
method	50	100	250	500	50	100	250	500
MOM	46.64	47.03	43.54	43.71	20.80	21.98	23.91	25.97
MF	72.80	69.76	65.24	64.17	21.22	20.51	20.26	20.42
HY	94.41	89.40	88.68	90.53	24.57	25.17	26.18	27.07
F	68.65	62.90	54.26	51.61	20.85	20.10	19.37	19.21
	n=2, p=0.75 sample size				n=5, p=0.5 sample size			
	50	100	250	500	50	100	250	500
MOM	78.32	92.73	95.23	78.08	37.38	38.65	37.22	38.74
MF	150.09	182.81	199.61	181.24	50.65	49.77	45.11	42.73
HY	160.04	190.60	214.52	206.26	54.51	52.98	49.89	50.75
F	132.52	176.79	190.64	164.41	50.51	48.82	45.49	43.41

Table 3.4
The mean number of iterations for all the methods when sampling from a 3-finite Poisson mixture

	p_1=0.4, p_2=0.3, λ_1=1, λ_2=5, λ_3=7 sample size					p_1=0.7, p_2=0.2, λ_1=1, λ_2=5, λ_3=10 sample size			
method	50	100	250	500		50	100	250	500
MOM	16.35	14.72	13.52	13.08		11.58	10.98	11.29	11.72
MF	18.49	17.44	17.61	17.50		13.82	13.14	13.43	13.68
HY	17.57	16.00	15.76	15.87		21.66	20.35	20.27	20.46
F	17.86	16.84	17.06	16.99		13.49	12.82	13.10	13.34
	p_1=0.3, p_2=0.4, λ_1=1, λ_2=2, λ_3=3 sample size					p_1=0.4, p_2=0.4, λ_1=1, λ_2=1.2, λ_3=5 sample size			
	50	100	250	500		50	100	250	500
MOM	78.56	81.64	83.91	67.89		25.64	19.91	16.60	14.34
MF	125.31	144.81	164.70	152.28		36.19	32.33	32.01	32.63
HY	130.49	143.97	155.33	139.17		44.63	39.70	37.57	36.94
F	112.21	141.70	159.26	151.46		35.91	32.59	32.17	32.68

Table 3.5
The proportion of times for which the initial estimates did not exist (based on 1000 replications), when sampling from a 2-finite Poisson mixture

	p_1=0.5, λ_1=1, λ_2=2 sample size					p_1=0.5, λ_1=1, λ_2=8 sample size			
method	50	100	250	500		50	100	250	500
TR	0	0	0	0		0	0	0	0
MOM	0.419	0.352	0.250	0.158		0.023	0.002	0	0
MF	0.282	0.162	0.048	0.008		0	0	0	0
HY	0.282	0.162	0.048	0.008		0.017	0	0	0
F	0	0	0	0		0	0	0	0
	p_1=0.8, λ_1=1, λ_2=5 sample size					p_1=0.1, λ_1=1, λ_2=8 sample size			
	50	100	250	500		50	100	250	500
TR	0	0	0	0		0	0	0	0
MOM	0.016	0	0	0		0.444	0.380	0.251	0.131
MF	0.001	0	0	0		0.036	0.004	0	0
HY	0.174	0.117	0.046	0.020		0.036	0.004	0	0
F	0	0	0	0		0	0	0	0

Table 3.6
The proportion of times for which the initial estimates did not exist (based on 1000 replications), when sampling from a negative binomial distribution

	n=1, p=0.5 sample size					n=2, p=0.25 sample size			
method	50	100	250	500		50	100	250	500
MOM	0.230	0.082	0.004	0		0.035	0.092	0.178	0.092
MF	0.007	0	0	0		0	0	0	0
HY	0.372	0.404	0.413	0.427		0.002	0.001	0	0
F	0	0	0	0		0	0	0	0
	n=2, p=0.75 sample size					n=5, p=0.5 sample size			
	50	100	250	500		50	100	250	500
MOM	0.575	0.401	0.194	0.075		0.028	0.001	0	0
MF	0.172	0.060	0.003	0		0.003	0	0	0
HY	0.259	0.114	0.016	0		0.003	0	0	0
F	0	0	0	0		0	0	0	0

Table 3.7
The proportion of times for which the initial estimates did not exist (based on 1000 replications), when sampling from a 3-finite Poisson mixture

	p_1=0.4, p_2=0.3, λ_1=1, λ_2=5, λ_3=7 sample size					p_1=0.7, p_2=0.2, λ_1=1, λ_2=5, λ_3=10 sample size			
method	50	100	250	500		50	100	250	500
MOM	0.049	0.010	0	0		0	0	0	0
MF	0	0	0	0		0	0	0	0
HY	0	0	0	0		0.601	0.712	0.850	0.946
F	0	0	0	0		0	0	0	0
	p_1=0.3, p_2=0.4, λ_1=1, λ_2=2, λ_3=3 sample size					p_1=0.4, p_2=0.4, λ_1=1, λ_2=1.2, λ_3=5 sample size			
	50	100	250	500		50	100	250	500
MOM	0.437	0.276	0.118	0.037		0.012	0.002	0	0
MF	0.133	0.047	0.003	0		0.001	0	0	0
HY	0.133	0.047	0.003	0		0.073	0.043	0.008	0
F	0	0	0	0		0	0	0	0

Table 3.8
The proportion of times the method converged to the global maximum when sampling from a 2-finite Poisson mixture

	p_1=0.5, λ_1=1, λ_2=2 sample size					p_1=0.5, λ_1=1, λ_2=8 sample size			
method	50	100	250	500		50	100	250	500
TR	0.70	0.66	0.64	0.58		1.00	1.00	1.00	1.00
MOM	0.79	0.76	0.76	0.72		1.00	1.00	1.00	1.00
MF	0.82	0.74	0.67	0.62		1.00	1.00	1.00	1.00
HY	0.83	0.72	0.65	0.58		1.00	1.00	1.00	1.00
F	0.72	0.68	0.63	0.54		1.00	1.00	1.00	1.00
	p_1=0.8, λ_1=1, λ_2=5 sample size					p_1=0.1, λ_1=1, λ_2=8 sample size			
	50	100	250	500		50	100	250	500
TR	0.98	0.99	1.00	1.00		0.94	0.96	0.99	1.00
MOM	0.98	0.99	1.00	1.00		0.91	0.95	0.99	1.00
MF	0.98	0.99	1.00	1.00		0.92	0.96	0.99	1.00
HY	0.97	0.99	1.00	1.00		0.91	0.96	0.99	1.00
F	0.98	0.99	1.00	1.00		0.90	0.96	0.99	1.00

Table 3.9
The proportion of times the method converged to the global maximum when sampling from a negative binomial distribution

	n=1, p=0.5 sample size					n=2, p=0.25 sample size			
method	50	100	250	500		50	100	250	500
MOM	0.95	0.96	0.97	0.98		0.95	0.95	0.96	0.98
MF	0.96	0.95	0.94	0.95		0.97	0.97	0.99	0.99
HY	0.93	0.90	0.88	0.91		0.96	0.96	0.98	0.99
F	0.97	0.96	0.95	0.96		0.97	0.97	0.99	0.99
	n=2, p=0.75 sample size					n=5, p=0.5 sample size			
	50	100	250	500		50	100	250	500
MOM	0.91	0.87	0.83	0.78		0.87	0.82	0.71	0.57
MF	0.88	0.84	0.72	0.64		0.90	0.85	0.76	0.73
HY	0.87	0.83	0.73	0.64		0.89	0.84	0.74	0.65
F	0.90	0.85	0.73	0.67		0.91	0.85	0.77	0.74

Table 3.10
The proportion of times the method converged to the global maximum when sampling from a 3-finite Poisson mixture

	p_1=0.4, p_2=0.3, λ_1=1, λ_2=5, λ_3=7					p_1=0.7, p_2=0.2, λ_1=1, λ_2=5, λ_3=10			
	sample size					sample size			
method	50	100	250	500		50	100	250	500
MOM	0.99	1.00	1.00	1.00		1.00	1.00	1.00	1.00
MF	1.00	1.00	1.00	1.00		1.00	1.00	1.00	1.00
HY	1.00	1.00	1.00	1.00		1.00	1.00	1.00	1.00
F	1.00	1.00	1.00	1.00		1.00	1.00	1.00	1.00
	p_1=0.3, p_2=0.4, λ_1=1, λ_2=2, λ_3=3					p_1=0.4, p_2=0.4, λ_1=1, λ_2=1.2, λ_3=5			
	sample size					sample size			
	50	100	250	500		50	100	250	500
MOM	0.80	0.77	0.73	0.70		0.99	0.99	1.00	1.00
MF	0.83	0.70	0.55	0.44		0.98	0.99	1.00	1.00
HY	0.81	0.69	0.55	0.47		0.97	1.00	1.00	1.00
F	0.85	0.74	0.58	0.49		0.98	1.00	1.00	1.00

Another interesting point is that when the components are well separated the algorithm converges very fast for all the sets of initial values. On the contrary, for mixtures with components close together the convergence is rather slow.

3.3.4 The Convergence of the EM Algorithm

The convergence of the general EM algorithm has been proved by Wu (1983) and Meilijson (1989). However it can be slow and one has to examine carefully if the global maximum has been reached. A problem encountered with real data is the choice of a criterion (stopping rule) for terminating the iterations.

A natural choice would be to stop the iterations when the increment in the loglikelihood between two successive iterations is smaller than some value, namely to stop if $\ell^{(i+1)} - \ell^{(i)} < tol$, where $\ell^{(i)}$ is the loglikelihood after i iterations and *tol* is some very small value used to show convergence. However, for such a criterion we have to take into account that the value of the loglikelihood depends on the sample size and hence, for small sample sizes, a small increase in absolute value may be important.

Another choice would be to stop iterating if $\frac{\ell^{(i)}}{\ell^{(i+1)}} > tol$, where *tol* is a number very close to 1, e.g. 0.99999. This stopping rule takes into account the relative improvement between 2 iterations. Agha and Ibrahim (1984) used this criterion.

Note that these criteria serve rather as lack of progress criteria than useful stopping criteria.

Everitt (1984b) proposed, in a normal mixture problem, to stop iterating when the Euclidean distance between the vectors of estimates between two iterations is less than 0.0001. This stopping criterion can have a substantial gain in computing time because at each iteration the time required for the computation of the loglikelihood is very long. This criterion can save time without great loss in accuracy.

A very interesting contribution to this problem is given by Finch *et al.* (1989). The authors estimated the probability that an iterative algorithm starting from different initial values fails to locate the global maximum. The lower the probability the more likely that the maximum is obtained. The idea is connected with the problem of estimating the unobserved species in ecology when we randomly sample from a population. The method proposed is very interesting since it provides us with an estimate of the probability that we have really found the maximum. All the other methods just examine whether the maximum is obtained, and thus any problem connected with the accuracy of the computer may destroy our confidence.

The results from Theorem 3.1 are useful at this point. We can search whether the obtained estimates satisfy the conditions to be ML estimates, given in Theorem 3.1. We have seen that the estimating equations satisfy conditions (a) and (b) of Theorem 3.1. Condition (c) verifies that the maximum has been obtained. However, this may be time consuming, since the examination as to whether the maximum has been obtained using condition (c) might need more computing time than that of an iteration itself. Thus, making more iteration might be more computationally efficient. In this case it would be useful to check for the convergence after a certain number of iteration and not after each iteration.

Another possible criterion would combine the improvement of the likelihood itself with some goodness of fit value. The underlying idea is that even though the likelihood is improving, the goodness of fit may be deteriorating after some iterations. Thus, if the likelihood has started to slow down we can stop iterating when the goodness of fit starts decreasing. The main problem is that we may stop too early in the sense that we will not find

the ML estimate, but a very close approximation. Note that, in fact, this is the case for every criterion proposed.

Another criterion is the Aitken acceleration applied to mixture models by Bohning *et al.* (1994) and McLachlan (1995). The EM algorithm converges only linearly and in practice the rate is often slow. This means that a large number of iterations is necessary to achieve parameter estimates of reasonable accuracy. Let $\ell^{\max}$ be the maximum of the loglikelihood. If the sequence $\ell^{(i)}$ converges linearly to $\ell^{\max}$ then

$\ell^{(i+1)} - \ell^{\max} \cong c(\ell^{(i)} - \ell^{\max})$ for all i and some c , 0 < c <1

where ≅ means that the equality holds for $i \to \infty$.

From this it is clear that if c is very close to 1, then a small increment in the loglikelihood need not mean that we are close to the maximum.

The improvement, known as Aitken acceleration, estimates the value of c at each iteration in order to predict the value at the limit and hence, to see if we are close to the value of the true maximum. So, we compute

$c_i = \dfrac{\ell^{(i+1)} - \ell^{(i)}}{\ell^{(i)} - \ell^{(i-1)}}$ and calculate the value

$\ell_\infty^{(i)} = \ell^{(i-1)} + \dfrac{1}{1-c_i}\left(\ell^{(i-1)} - \ell^{(i)}\right)$. Iterations stop when $\left|\ell_\infty^{(i)} - \ell_\infty^{(i-1)}\right| < tol$, where *tol* has some fixed value. Note that the value of $\ell_\infty^{(i)}$ can be used as a prediction of the true maximum loglikelihood value. This can be useful in some cases, e.g. when we want to perform a likelihood ratio test.

Jones and McLachlan (1992) used a similar device for an EM algorithm for finite mixtures for grouped data.

The EÌ converges linearly to the maximum, while other numerical methods converge quadratically. These include the Newton-Raphson method. On the contrary, such numerical methods require good initial values near the maximum in order to converge. So, a good strategy might be to start with EM iterations to get nearer the maximum and then continue with Newton-Raphson iterations to locate it (see Aitkin and Aitkin, 1996).

Example 3.1 (continued) In order to show how important the convergence criterion is, we run the EM algorithm for the data of Example 3.1, starting from the same point, but using two different stopping rules. The two rules were to stop iterating when $\dfrac{\ell^{(i)}}{\ell^{(i+1)}} > tol$, with

$tol = 9.9 \times 10^{-5}$ for the first criterion and $tol = 9.9 \times 10^{-15}$ for the second. Table 3.11 contain the results. The starting values were p_1=0.5, $\ddot{e}_1$ =1 and $\ddot{e}_2$ =2.

Table 3.11
The estimates derived via the EM algorithm for the data in Table 3.1, using two different stopping rules

criterion	p_1	$\ddot{e}_1$	$\ddot{e}_2$	number of iterations	loglikelihood
$tol = 9.9 \times 10^{-5}$	0.7197	1.5652	3.9768	39	-274.1617
$tol = 9.9 \times 10^{-15}$	0.6727	1.4888	3.7885	334	-274.1226

Table 3.11 reveals a very important problem. Using the second stopping rule we achieved an increase of the loglikelihood by only 0.0002%. But the change in the parameter values is very large. In fact the mixing proportion changed almost 7%, while each of the component parameters changed almost 5%. This example illustrates that a negligible change in the loglikelihood can lead to a significant change of the estimates. The standard errors for parameters p_1 , $\ddot{e}_1$, $\ddot{e}_2$ were found to be equal to 0.178, 0.304 and 0.715 respectively. Using Aitken acceleration we verified that the maximum has been obtained. One can see that the changes on the parameters are smaller relative to their standard errors. Lindsay (1995, page 131) discussed a stopping rule based on both likelihood issues and the standard errors of the parameters.

3.3.5 Applications

The EM algorithm is widely accepted as a method which can be used easily in mixture models. This is the reason why it is very popular among researchers and for the majority of cases of ML estimation in mixture models the EM algorithm is used. For a wealth of applications in mixture models one can refer to the book of Titterington *et al.* (1985).

Leroux and Puterman (1992) extended the method to Markov dependent Poisson process. In this case instead of a simple parameter for each component we have to estimate the stochastic matrix as well.

Some other applications involve random coefficient regression models, Mallet (1986), outlier identification Aitkin and Wilson (1980), grouped or truncated data, Mclaren *et al.* (1986), McLachlan and Jones (1988), longitudinal data, Dietz and Bohning (1994) latent class models Aitkin *et al.* (1981). De Vaux (1989) treated the case of mixtures of linear regression, known also as switching regressions, applying the EM algorithm. Poisson mixtures are treated

in Symons *et al.* (1983) for a cluster application and in Gibbons *et al.* (1990). A number of recent applications can be found in Lindsay and Roeder (1995).

Another interesting application of the EM algorithm for mixtures can be found in Jorgensen (1990). He discussed influence-based diagnostics for finite mixture models and he described the idea of measuring the influence of an observation by calculating the difference of the loglikelihood if this observation is removed. Since this is time consuming for mixture models, because of the difficulty in obtaining the ML estimates, he proposed to use the so-called one step influence by substituting the loglikelihood of the new sample by the loglikelihood obtained after one EM iteration, starting from the ML estimate of the full data set.

3.4 Variants of the EM Algorithm and Related Algorithms

3.4.1 Variants of the EM

The EM algorithm can serve as a powerful tool for ML estimation for all cases when the model can be written as a mixture model. The general iterative scheme can be used in all the circumstances when some of the parameters involved are known. As a general example one can refer to the case of normal mixtures. The more general case involves the estimation of both λ_i's and σ_i's, while in some situations we may consider the variances known and thus only the estimation of λ_i 's is necessary. In this case the general iterative scheme applies keeping the rest of the known parameters fixed (i.e. we do not need to update our estimates at each iteration for these parameters). As far as the Poisson case is concerned we may use variants of the EM in order to estimate the parameters of some related distributions. Three cases are given.

Case 1: Distributions with Added Zeroes

Example 3.2 Guillen and Artis (1992) presented the data set of Table 3.12. The random variable X represents the number of defaulted instalments, i.e. the number of times the client did not pay the money as it was agreed when credit was granted. The examination of such data sets is crucial since any financial institution would want to know how many of its clients will not pay their instalments, especially when new contracts are signed.

Table 3.12
The number of defaulted instalments

x	frequency	x	frequency	x	frequency	x	frequency
0	3002	9	53	18	8	27	0
1	502	10	41	19	6	28	1
2	187	11	28	20	3	29	1
3	138	12	34	21	0	30	1
4	233	13	10	22	1	31	1
5	160	14	13	23	0	32	0
6	107	15	11	24	1	33	0
7	80	16	4	25	0	34	1
8	59	17	5	26	0		

The financial institution is interested in classifying the clients as 'good' and 'bad', i.e. those who pay their instalments and those whose instalments are defaulted. For the 'good' clients, $X=0$, i.e. there are no defaulted instalments. Thus, the entire population consists of at least two groups, the first group is the group of 'good' clients which has certainly the value x=0, while for every other group the random variable X follows a $Po(\ddot{e})$ distribution with some group specific value of the parameter $\ddot{e}$. We suppose that there are two groups of 'bad' clients. This model give rise to a 2-finite Poisson mixture with added zeros.

Johnson *et al.* (1992) described the case of Poisson with added zeros or inflated Poisson distribution. This is the distribution where some observations with value equal to zero have been added to the simple Poisson distribution. This distribution can be considered as a mixed Poisson with the first component having a Poisson distribution with parameter equal to 0 (this is a distribution degenerate at 0).

A 2-finite Poisson mixture with added zeroes can be regarded as a 3-finite Poisson mixture whose first component has a parameter equal to 0. So, the standard EM algorithm is applicable, but at each step we do not need to update the value of the parameter $\ddot{e}_1$.

We applied the EM algorithm for the data in Table 3.12. Table 3.13 contains the values of the parameters of the fitted distribution.

Table 3.13
The fitted 3-finite Poisson mixture for the data in Table 3.12

	mixing proportion	component parameter
1st	0.605 (0.0116)	0
2nd	0.276 (0.0132)	2.078 (0.262)
3rd	0.119	8.506 (0.698)
	loglikelihood=-7211.51	

According to the specified model the proportion of 'good' clients is 0.605. The standard errors of the parameters are shown in the parentheses.

Fong and Yip (1993) described the EM algorithm for several discrete distributions with added zeroes. Their algorithm is the standard EM algorithm for mixtures described above.

This scheme can be generalised by fixing the value of one (or more) component parameters. Our example treated the case when the parameter was fixed to have a 0 value. The general EM algorithm applies, but at each iteration we do not update the values of the parameters. Hence, the EM algorithm can be used in cases when only the mixing proportions have to be estimated.

Case 2: Discrete Mixing Distribution.

Let us now restrict our attention to the case where the mixing distribution is finite but the parameters $\ddot{e}_i$ take integer values 1,2,. . ., k . We may allow $\underline{\ddot{e}_i}$ to take the value 0 too. In this case the estimation can be made via the EM ignoring the step for the estimation of $\ddot{e}_i$, $i=1,2,\ldots,k$ and using the fixed values at each iteration. This corresponds to the case of a discrete mixing distribution. Note that we may use this estimated mixing distribution for goodness of fit purposes. For example, the Neyman distribution is known to be a mixed Poisson distribution with the Poisson itself as the mixing distribution (see, e.g., Douglas, 1980). Thus for estimating a discrete mixing distribution we may check if this distribution is the Poisson distribution or not. The identifiability of Poisson mixtures guarantees the equivalence of the two tests. Tests for the Poisson distribution are more common and well examined than tests for the goodness of fit of the Neyman distribution. Finding a mechanism which reduces a hypothesis testing procedure for the mixed Poisson distribution to a hypothesis testing procedure for the form of the mixing distribution remains an open problem.

Example 3.1 (continued) Consider the data of Example 3.1 and assume that the mixing distribution is discrete. Under this assumption, we applied the EM algorithm, allowing the value 0 for the mixing distribution, and we estimated the discrete mixing distribution (recall that the Poisson distribution with mean 0 is the degenerate distribution at 0). Note that we assume that the mixing distribution takes values 0,1,..,9. As the results in Table 3.14 show,

this restriction does not play any role in the procedure and simply facilitates the estimation procedure. Table 3.14 contains the estimated discrete mixing distribution.

Table 3.14
The estimated discrete mixing distribution

x	0	1	2	3	4	5	6
Probability	0	0.2549	0.4461	0.1517	0.0994	0.0463	0.0013

Figure 3.1 depicts the estimated discrete mixing distribution along with the estimated Poisson distribution with parameter equal to 2.241 (i.e. equal to the sample mean). Clearly the estimated distribution differs from the Poisson distribution and thus one may conclude that the Neyman distribution, which is the mixed Poisson distribution obtained using a Poisson distribution as the mixing distribution, is not appropriate to describe this data set.

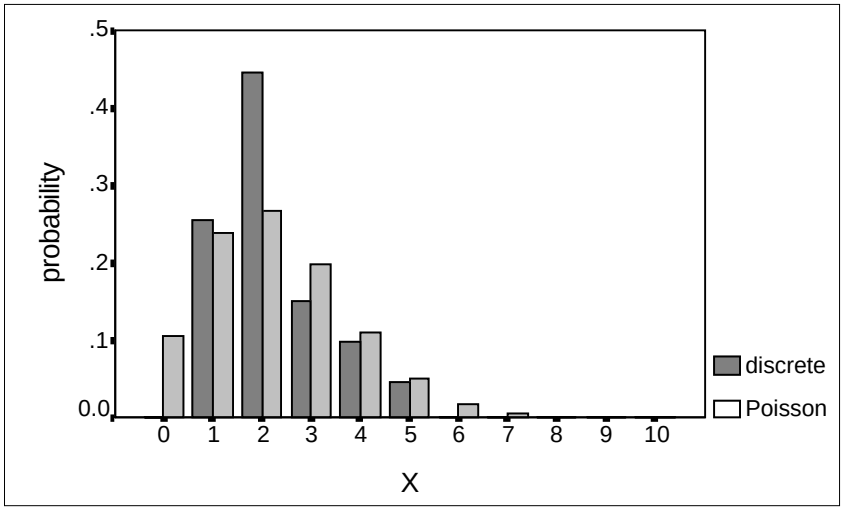


Figure 3.1 The estimated discrete mixing distribution for the data of Example 3.1 and the Poisson distribution with ë=2.241.

Case 3: Finite Mixtures with Different Components

We can generalise the EM algorithm, to cover cases where the components of the mixture follow different distributions. Only slight modifications are needed in the M-step, where the ML estimate of each component distribution must be obtained. We will present a simple example.

Example 3.1 (continued) Instead of assuming a 2-finite Poisson mixture we assume that our data come from a mixture of a Poisson distribution with a geometric distribution as defined by (2.24). Then we assume that the random variable X follows a discrete distribution with probability function given by

$$P(x) = p\frac{e^{-\lambda}\lambda^x}{x!} + (1-p)\left(\frac{a}{1+a}\right)\left(\frac{1}{1+a}\right)^x, \qquad (3.21)$$

where *x=0, 1, 2, . . . , á, ë*>0, and $0 \le p \le 1$. If *p=1* we obtain the simple Poisson distribution while if p=0 we obtain the simple Geometric distribution.

The EM algorithm can be described as :

Step 1 (E-step) : Given the current values for λ^{old}, a^{old} and p^{old} we calculate the probability w_{ij} that the observation X belongs to the j subpopulation (j=1,2) after observing it (namely the posterior probability of belonging in the j subpopulation). Note that the first subpopulation follows a Poisson distribution while the second follows a geometric distribution. These posterior probabilities are calculated as:

$$w_{i1} = \frac{p^{old} f_1(x_i|\lambda^{old})}{P(x_i)} \qquad \text{and} \qquad w_{i2} = \frac{\left(1-p^{old}\right) f_2(x_i|a^{old})}{P(x_i)} \qquad (3.22)$$

where *P(x)* is calculated by (3.21), and $f_1(x|\lambda)$ and $f_2(x|a)$ are the probability functions of the Poisson and the Geometric distributions respectively.

Step 2 (M-step) : Calculate the new estimates as

$$\lambda^{new} = \frac{\sum_{i=1}^{n} w_{i1}x_i}{\sum_{i=1}^{n} w_{i1}}, \quad a^{new} = \frac{\sum_{i=1}^{n} w_{i2}}{\sum_{i=1}^{n} w_{i2}x_i} \quad \text{and} \quad p^{new} = \frac{\sum_{i=1}^{n} w_{i1}}{n} \qquad (3.23)$$

Step 3 : Check if some condition is satisfied in order to terminate the iterations, otherwise go back to step 1, using the currently estimated values.

Applying this scheme to the data of Example 3.1 we derived the following estimates (with their jacknife standard errors in parentheses): p=0.707 (0.121), ë=2.269 (0.283) and á=0.459 (0.113). The maximised loglikelihood is -276.1402, which shows that the 2-finite Poisson mixture model is more plausible.

Note also that mixtures of the form given in (3.21) can be used for testing the assumption that data come from the Poisson distribution against the alternative that data come from the geometric distribution. So the hypothesis testing reduces to testing if p=1. Durairajan and Kale (1979, 1982) described hypothesis testing for the mixing proportion. A similar idea for goodness of fit tests is described in Rudas *et al.* (1994).

Rachev and SenGupta (1994) describe a finite mixture of Laplace-Weibull distributions for modelling the price changes in the stock market. Scallan (1992) described a finite Normal-Laplace distribution for modelling the wind shear. Al-Hussaini and Abd-El-Hakim (1990) proposed an inverse Gaussian-Weibull mixture for reliability applications.. Clearly the EM algorithm is a useful tool for describing mixtures with different components. Moreover such models can be used for model selection purposes.

3.4.2 Related Algorithms

a) The ECM Algorithm

Another class of algorithms, for Poisson mixtures, very similar to the EM is the expectation-conditional maximisation (ECM) algorithm. Meng and Rubin (1993) discussed this class of EM algorithms. The idea is that in many cases complete data ML estimation is relatively simple conditional on some function of the parameters being estimated. In other words, when we cannot maximise the likelihood directly because of its complexity we can maximise it with respect to one parameter keeping the other parameters fixed (i.e. we maximise the likelihood with respect to another parameter conditionally on the other parameters and so on). Meng and Rubin (1993) showed that the ECM algorithm has the same properties as the EM algorithm and they discussed further the ECM algorithm when the maximisation step of the EM algorithm is replaced by several ECM steps.

In each step we estimate the values of p_i ´s for the given values of ë$_i$ ´s and then we try to find the $\ddot{e}_i$ ´s which maximise the likelihood with the given p_i ´s . The concavity of the functions needed to be maximised can be verified easily. Because of the fact that the method is a special case of the EM algorithm its convergence is ensured. Of course, it is not known whether the algorithm converges to the global maximum (and not to a local maximum). So, we have to try several initial values or we have to test if the global maximum is obtained. However, the ECM algorithm shows no clear improvement over the EM algorithm.

b) The SEM Algorithm

Celeux and Diebolt (1985,1992) discussed a stochastic version of the EM algorithm named SEM (Stochastic EM). The idea is to replace the closed form expression for the E-step with a stochastic step, by simulating the expected values from a multinomial distribution with parameter n (the sample size) and probabilities equal to the posterior probabilities for given

values of the parameters calculated as in (3.16). These posterior probabilities are those calculated at the E-step of the EM algorithm. The dual idea for this step is primarily to avoid small sample size difficulties and secondly to use the fact that the variability from the stochastic step accounts for the variability of the data. This method works well for moderate or large sample sizes, and overcomes most of the limitations of the pure EM algorithm. For example, the simple EM algorithm proceeds towards the nearest local maxima. The stochasticity of the SEM algorithm can help in locating maxima other than the nearest local ones. On the other hand, it converges only in distribution and usually it takes more computational time. It must be pointed out that, since the SEM provide only convergence in distribution, it does not lead to point estimates, as the EM algorithm does. However, a point estimate can be obtained by averaging a sufficient number of successive estimates after the procedure has reached stationarity. See also Diebolt and Celeux (1993) for asymptotic properties of the SEM algorithm for mixtures and Chauveau (1995) for applying the SEM algorithm in finite mixtures with censored data. Diebolt and Ip (1996) provide a review on the SEM algorithm.

Example 3.1 (continued)

We applied the SEM algorithm to the dataset concerning the number of crimes in Greece, given in Table 3.1. After 100 iterations to reach the maximum 1000 values were considered for representing the distribution of each parameter. The means of these distributions representing the point estimates of the parameters were p_1=0.6784 (0.134), $\ddot{e}_1$=1.5022 (0.234) and $\ddot{e}_2$=3.926 (0.566).

Figure 3.2 depicts the estimated distributions of the parameters via the SEM algorithm (smoothed via kernel density methods).

Figure 3.2a

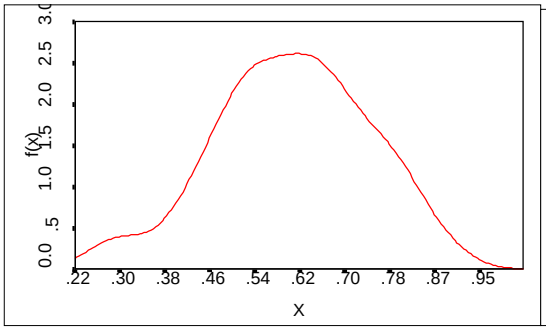


Figure 3.2b

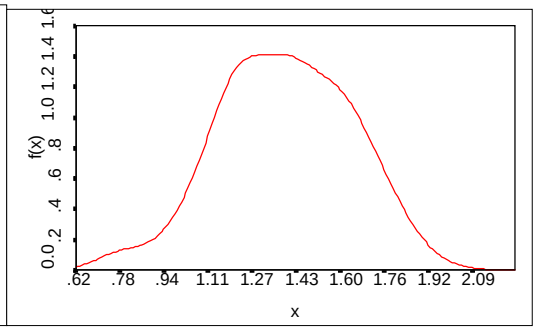


Figure 3.2c

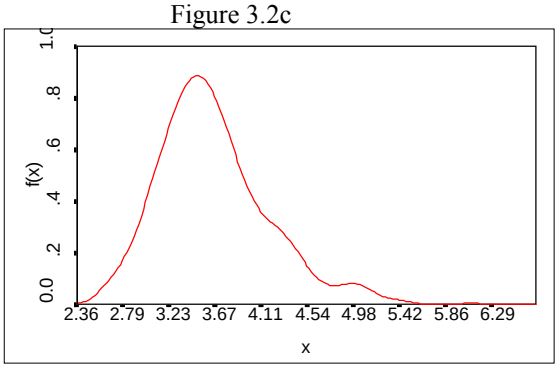


Figure 3.2 The distributions of the 3 parameters p_1, $\ddot{e}_1$ and $\ddot{e}_2$ (figures a, b and c respectively) derived via the SEM algorithm.

c) The SAEM algorithm

The SAEM algorithm (Stochastic Approximation EM algorithm), is a compromise between the EM algorithm and the SEM algorithm. This algorithm also uses the S-step but the probabilities used in the M-step are a weighted average of the probabilities of the E-step and those of the S-step. The weight in the i-th iteration has a value γ_i , where γ_0 =1 and $\{\gamma_n\}$ is a sequence of positive real numbers in (0,1) decreasing to zero at a sufficiently slow rate. Symbolically, SAEM = (1-γ_n) EM+ γ_n SEM. Hence the SAEM algorithm starts of a pure

SEM algorithm and evolves to a pure EM algorithm eventually. The convergence is almost sure. SAEM works better for small sample sizes.

Celeux *et al.* (1995) described all of the above mentioned stochastic versions of the EM algorithm.

d) Other algorithms

A variant of the EM algorithm, named the Monte Carlo EM algorithm, has also been proposed (see Wei and Tanner (1990)). The idea is to calculate the E-step via simulation of a large number of replications the quantity for which we want to calculate its expectation, and then to estimate its expectation with the mean of these replications. This algorithm is very useful when the E-step cannot be performed in closed form.

Aitkin and Rubin (1985) described another variant of the standard EM algorithm. They proposed to use a prior distribution for the mixing proportions. Therefore at each E-step one should integrate out the mixing proportions under this prior. The aim of the authors was to ensure that the estimates always lie in the interior of the parameter space not approaching 0 or 1. This assumption is useful when one wants to apply the standard likelihood ratio test.

Nychka (1990) proposed the addition of a smoothing step after the maximisation step. This idea can be transferred to the finite mixture case by trying to smooth the estimates from the i-th iteration for the mixing parameters. This may be interesting if we are trying to estimate a great number of parameters (large k). Nychka (1990) showed that his proposal is similar with a penalised ML estimation method and the smoothing step is the application of the penalty.

Another algorithm which uses the EM algorithm was proposed by Leonard *et al.* (1994). They proposed to start from an equal weighted mixture, namely assuming that each component has the same mixing proportion. Then, using the iterative scheme of the EM they estimate the parameters of interest only, and if these parameters are very close together they replace them with their mean and the new mixing proportion is the number of components clustered divided by the initial number. This enables one to estimate the number of components too. A similar idea can be found in Bohning *et al.* (1992). In this case the mixing proportions were estimated, and then components close together or components with very small mixing proportion were clustered. Leonard *et al.* (1994) proposed not to estimate the mixing proportions via the EM but with a perturbing search among the established clusters. The only disadvantage is the complexity of this search, because the search needs to subtract a

value of each component and to add it to another one, looking for whether the likelihood is increased.

The increased complexity in this last step is compensated by the economy in the first step when the mixing proportions were not estimated. The number of initial components must be large enough. The authors proposed to start the search from a number equal to the sample size.

De Vaux and Krieger (1990) described a method to robustify the EM algorithm for normal mixtures. In fact, their approach does not lead to ML estimates but, as the authors showed, their proposal reduces the risk of inconsistent estimates due to the presence of some outliers. The idea for robustification is to replace the used measures either at the E-step or at the M-step or at both steps with robust counterparts. For example, in normal mixtures the authors proposed to use at the M-step, the median instead of the mean and the mean absolute deviation instead of the variance. For the E-step they proposed to use more heavily tailed distributions like the Student or the double exponential. From the simulation results reported such an approach may preserve the estimates from the influence of outliers. Clearly, the estimators are not ML estimators but some kind of distance estimators.

3.5 Improving the EM Algorithm for Mixtures: A New Method

In section 3.2.2 we saw that the ML estimates for finite mixtures of the one-parameter exponential family satisfy the first moment equation. This result can be used to simplify the ML estimation for finite mixtures and more specifically to improve the EM algorithm.

The EM algorithm, despite its disadvantages, is the commonly used method for ML estimation for finite mixtures. Improvements of the EM algorithm for finite mixtures have been proposed in three different directions.

- Bohning *et al.* (1994) proposed methods that can easier detect the convergence of the algorithm and thus saving iterations.
- Fruman and Lindsay (1994) recommended the use of efficient initial values, namely the use of the moment estimates as initial values for the EM algorithm.
- Aitkin and Aitkin (1996) proposed that we can speed up the convergence by alternating the EM iterations with Gauss-Newton iterations (see also Lange, 1995, Jamshidian and Jennrich, 1997)

Our results of section 3.2.2 can serve as a basis for improving the EM algorithm for finite mixtures. At each step we do not have to calculate all the 2k-1 parameters. It suffices to calculate 2k-2 of them while the rest can be easily calculated from the moment equation at a low cost. This can save much of the computational time needed for each iteration because the updated estimate is a sum with a large number of summands as it can be seen from (3.17)-(3.18). Note that our approach can be combined with the above mentioned methods in order to maximise the gain in computing time.

This improvement applies to all members of the exponential family defined in (3.7) for which the results of section 3.2.2 hold. The gain in computing time is considerable as shown in Tables 3.15-3.17 for small values of k and for finite mixtures from the Poisson and the Normal distributions. If we look at the iterative scheme described in (3.16)-(3.18), we can see that we can avoid calculating $\grave{\imath}(\grave{e}_k)$ (i.e. the parameter of the k-th component) and this is equivalent to reducing the calculations involved for obtaining the new parameters by almost the 1/(2k-1). In fact, the gain is less because of the cost for some additional calculations at each iteration. It is also interesting that the gain is expected to be larger in the case of discrete distributions, like the Poisson or the binomial distributions. This is so because, for discrete distributions, we can avoid exhausting summations by multiplying with the observed frequencies.

We give two examples to show the gain in time using our method.

Example 3.3 Finite Poisson mixtures

Consider the case of finite Poisson mixtures. In order to examine the gain we carried out a small simulation comparison. For k=2 we simulated 100 samples of given sample size n (n=50, 100, 250, 500) for each distribution with parameter vectors (p_1= p , $\ddot{e}_1$ = 1, $\ddot{e}_2$). We calculated the necessary time for ML estimation via the EM algorithm using both the general EM algorithm and the improved EM algorithm discussed above. Each entry in Table 3.15 represents the time that the improved EM algorithm takes relative to that taken by the standard EM algorithm, i.e. the time of the improved EM, divided by the time of the standard EM algorithm. From both the numerator and the denominator we have subtracted the time spent for simulating the samples. We tried to minimise the computing time for some auxiliary procedures like the terminating conditions. For each sample we stopped running the algorithm after 50 iterations. All the calculations were carried out in a PC with Pentium microprocessor (120 Hz). The results of Table 3.15 clearly show that we can save almost 20% of the computing time for k=2.

Table 3.15
Times for the improved EM relative to the standard EM for 2-finite Poisson mixtures (k=2)

	p_1 =0.25			p_1 =0.50			p_1 =0.75		
$ë_2$	2	5	10	2	5	10	2	5	10
n									
50	0.822	0.803	0.797	0.821	0.800	0.799	0.831	0.815	0.806
100	0.823	0.800	0.792	0.812	0.799	0.797	0.789	0.808	0.800
250	0.809	0.795	0.792	0.813	0.796	0.795	0.819	0.803	0.794
500	0.813	0.793	0.791	0.812	0.752	0.794	0.809	0.820	0.793

Table 3.16 contains the results for k=3 (3-finite Poisson mixture). The vectors of parameters were (p_1 , p_2 = 0.3, $ë_1$=1, $ë_2$=2, $ë_3$). For each distribution 100 samples of given sample size n (n=50,100,250,500) were simulated and the times required for both methods were recorded. The entries are again the time for the improved EM, divided by the time for the standard EM. We can also see an improvement on the required computational time near 15%.

Table 3.16
Times for the improved EM relative to the standard EM for 3-finite Poisson mixtures (k=3)

	p_1 =0.25			p_1 =0.50			p_1 =0.75		
$ë_3$	3	5	10	3	5	10	3	5	10
n									
50	0.869	0.859	0.853	0.868	0.863	0.857	0.875	0.870	0.866
100	0.866	0.860	0.853	0.866	0.861	0.854	0.871	0.862	0.858
250	0.863	0.857	0.851	0.862	0.857	0.851	0.863	0.859	0.854
500	0.862	0.856	0.850	0.859	0.855	0.850	0.863	0.858	0.851

Example 3.4: Finite Normal mixtures

Behboodian (1970) showed that for the case of normal mixtures with different variances, the second moment equation is also satisfied, i.e. the variance from the ML estimates is the same as the sample variance. So, in the case of normal mixtures, at each EM iteration we can simplify the estimation of 2 parameters. The total number of parameters to be estimated is 3k-1. We have only to calculate the 3k-3 parameters, while the remaining 2

parameters can easily be obtained from equating the first two moments. Thus we reduce the effort almost by a factor of 2/(3k-1). In practice the gain is less than this factor because of the cost of some additional calculations in each iteration, but it is still helpful. For k=2, we simulated 100 samples from several 2-finite normal mixtures. The gain is near 30% (we estimate 3 instead of 5 parameters) as we can see in Table 3.17, for selected parameter vectors $\grave{e} = (p_1, \grave{i}_1, \grave{i}_2, \acute{o}_1^2, \acute{o}_2^2)$ and varying sample sizes. The entries of Table 3.17 are again the ratios of computing times required by the improved EM divided by the corresponding computing times required by the standard EM algorithm for the same samples. Again, we tried to minimise any auxiliary calculations, and the comments from the Poisson case apply too.

Table 3.17
Times for the improved EM relative to the standard EM for 2-finite Normal mixtures (k=2)

vector of parameters $(p_1, \grave{i}_1, \grave{i}_2, \acute{o}_1^2, \acute{o}_2^2)$	sample size			
	n=50	n=100	n=250	n=500
(0.25, 0, -1, 1, 2)	0.707	0.731	0.720	0.724
(0.25, 0, 1, 1, 2)	0.731	0.706	0.747	0.693
(0.5, 0, -1, 1, 2)	0.722	0.725	0.729	0.730
(0.5, 0, 1, 1, 2)	0.719	0.728	0.725	0.722
(0.75, 0, -1, 1, 2)	0.711	0.717	0.723	0.708
(0.75, 0, 1, 1, 2)	0.729	0.732	0.735	0.727
(0.25, 0, -1, 1, 5)	0.728	0.730	0.718	0.726
(0.25, 0, 1, 1, 5)	0.726	0.731	0.736	0.731
(0.5, 0, -1, 1, 5)	0.729	0.733	0.729	0.734
(0.5, 0, 1, 1, 5)	0.731	0.728	0.727	0.730
(0.75, 0, -1, 1, 5)	0.724	0.729	0.731	0.710
(0.75, 0, 1, 1, 5)	0.720	0.724	0.717	0.721

These two examples reveal that a substantial improvement in computing time can be achieved by using this simple relation. It is important to note that similar improvements can be made for any other iterative algorithm.

3.6 M2 Type Samples: Maximum Likelihood Estimation

3.6.1 An EM Algorithm for Maximum Likelihood Estimation for M2 Type Samples from Finite Poisson Mixtures

Let us now examine the case when some additional information is available. This is the case when we have an M1 or an M2 type of sample. We will treat the case of M2 type

samples since this is more common in practice. This is the case where, for a certain proportion of our observations, we know the subpopulations to which they belong.

Suppose for example that we are examining accident data from an insurance company. Because of the fact that some of the records of the company are incomplete the age of the driver has not been recorded for all records. Fortunately, for a proportion, say á, of them the age is recorded. This sample is an M2 type sample since for some observations we have additional information.

Hosmer (1973a) examined the ML estimation method for normal mixtures under the three types of samples described in the previous section, using the EM algorithm. He concluded that a very small proportion of known data over the whole sample can lead to considerable gain in efficiency of the estimators.

Let us describe the case in Poisson mixtures when there are some cases for which we know the subpopulation to which they belong. Suppose that we have n observations from the mixture and m more observations of which m_j come from the j-th subpopulation, with $\sum_{j=1}^{k} m_j = m$. Then we have a total sample size of N=n+m observations and for only m of them we know their true subpopulation. For simplicity suppose that we have rearranged our sample so that the first n observations are those without additional information followed by the observations with additional information. We introduce the notation $M_j = \sum_{i=1}^{j} m_i$, i.e. M_j represents the number of observations of known origin which belong to the j-th component j=1, . . ., k. With this notation, our M2 type sample is:

$$(X_1,\ldots,X_n,X_{n+1},\ldots,X_{n+M_1},X_{n+M_1+1},\ldots,X_{n+M_2},\ldots,X_{n+M_{k-1}+1},\ldots,X_{n+m})$$

Now the likelihood of the sample is given by

$$L=\left\{\prod_{i=1}^{n}\left(\sum_{j=1}^{k}p_j f(x_i|\lambda_j)\right)\right\}*\binom{m}{m_1 m_2\ldots m_k}p_1^{m_1}p_2^{m_2}\ldots p_k^{m_k}*\prod_{j=1}^{k}\prod_{i=n+M_{j-1}+1}^{n+M_j}f(x_i|\lambda_j) \tag{3.24}$$

We have factorised the likelihood in three factors.

- The first factor is the contribution of the observations of unknown origin.
- The second factor is the probability of taking such a partition of the known observations, i.e. the probability of choosing m persons. We obtain m_j ,j=1, . . ., k from the j-th subpopulation. This is a multinomial probability.

•The third factor is the contribution of the observations of known origin.

The loglikelihood ℓ is written as

$$\ell = \sum_{i=1}^{n} \ln\left(\sum_{j=1}^{k} p_j f(x_i|\lambda_j) \right) + \ln\left(\binom{m}{m_1 m_2 \ldots m_k} p_1^{m_1} p_2^{m_2} \ldots p_k^{m_k} \right) + \sum_{j=1}^{k} \sum_{i=n+M_{j-1}+1}^{n+M_j} \ln f(x_i|\lambda_j) \quad (3.25)$$

Taking the partial derivatives of the loglikelihood given in (3.25) and equating them to 0, one can theoretically find the ML estimates. However analytic solution of the above problem is not straightforward and numerical techniques are necessary. We can use the EM algorithm to derive the ML estimates. We will demonstrate now that this approach is very similar to the one described for the simple case via the EM algorithm.

The iterative EM algorithm is almost the same as the one given in (3.16)-(3.18) for the simple sample type. The only difference is that now we do not have to estimate the posterior probabilities of belonging to the j-th class, given in (3.16) for the observations with known origin. The reason is that these values are known, i.e. we do not need to consider them as 'missing', so as to try to estimate them at the E-step.

Then (3.16) can be rewritten for the case of the EM algorithm for M2 type of sample as:

$$w_{ij} = \frac{p_j f(x_i|\lambda_j)}{f_P(x_i)} \qquad \text{for } i=1, 2, \ldots, n \text{ and } j=1, \ldots, k \quad \text{and}$$

w_{ij} = 1 if the i-th observation belongs to the j subpopulation and 0 otherwise, for $i = n+1, \ldots, n+m$.

Then at step 2 (the M-step) we calculate again the new parameters as

Step 2 (M-step) Calculate $\lambda_j = \frac{\sum_{i=1}^{N} w_{ij} x_i}{\sum_{i=1}^{N} w_{ij}}$ and $p_j = \frac{\sum_{i=1}^{N} w_{ij}}{N}$ for $j=1, 2, \ldots, k$

Step 2 is the same as the maximisation step of the EM algorithm for simple samples given in (3.16)-(3.17). The only difference is that now we do not try to estimate the posterior probabilities w_{ij} for the cases with known origin. Instead, we use the supplementary information setting this probability equal to 1 if the i-th observation belongs to the j-th population and 0 otherwise. This constitutes the first instance of an illustration for the case of Poisson mixtures from M2 type of sample in the literature.

We follow the matrix representation of the EM algorithm , given in (3.19) and (3.20) with the only difference in the construction of the **W** matrix. It is an N x k matrix of the form

$$\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1k} \\ . & . & \dots & . \\ w_{n1} & w_{n2} & \dots & w_{nk} \\ 1_{m_1} & 0_{m_1} & \dots & 0_{m_1} \\ 0_{m_2} & 1_{m_2} & \dots & 0_{m_2} \\ . & . & \dots & . \\ . & . & \dots & . \\ 0_{m_k} & 0_{m_k} & \dots & 1_{m_k} \end{bmatrix},$$

where $\mathbf{1}_n$ is a vector with all its n elements equal to 1 and $\mathbf{0}_n$ is the vector with all its elements equal to 0. Equations (3.19) and (3.20) still hold for obtaining the estimates of the parameters.

3.6.2 A Simulation Comparison

Intuitively, we expect that an M2 type sample is preferable to the typical sample with no additional information because of the complementary information contained in it. We carried out a simulation experiment to examine the gain from this additional information.

From some 2-finite Poisson distributions we simulated 10000 samples of given sample size n (n=50,100,500). For simulating from the 2-finite Poisson distributions, we used an algorithm which firstly chooses the component from which it will simulate a Poisson variate and then it simulates this value. This enables us to record the true value of the subpopulation of the observation. For a proportion á of the observations (á = 0.1, 0.2, 0.3) this true subpopulation value was recorded, and the estimated parameters using the EM algorithm for both the sample with no additional information and for the sample with additional information were calculated.

Then the standard errors of the parameters as well as the mean squared errors were calculated from the 10000 values for the parameters, for each sample type. In Tables 3.18-3.19 we can see the relative efficiencies and the relative mean squared errors for the parameters. The relative efficiency was calculated as the ratio of the standard error of the parameter if no additional information is available, divided by the standard error of the parameter using the additional information. So, entries larger than 1 favour the case with additional information. Similarly the relative mean squared error was calculated as the ratio of the mean squared error of the parameter if no additional information is available, divided by the mean squared error

of the parameter using the additional information. Again, entries larger than 1 favour the case with additional information.

Table 3.18
Relative Efficiencies. The entries are the ratios of the standard deviations of the parameter estimates when no supplementary information is available to those when the true subpopulation is known for a fraction á of the data

		α								
Parameter Vector		0.1			0.2			0.3		
	n	p_1	$ë_1$	$ë_2$	p_1	$ë_1$	$ë_2$	p_1	$ë_1$	$ë_2$
(0.2,1,2)	50	1.19	1.37	1.56	1.31	1.33	1.85	1.30	1.36	1.85
	100	1.54	1.80	1.89	1.94	1.87	2.30	1.71	1.87	1.93
	500	4.43	3.78	2.94	4.44	3.66	3.23	3.51	3.40	2.68
(0.5,1,2)	50	1.20	2.02	1.60	1.31	2.08	1.63	1.32	2.03	1.79
	100	2.08	2.61	2.07	1.99	2.68	1.85	1.99	2.73	1.92
	500	5.59	5.06	3.58	5.31	4.91	3.57	5.39	4.68	3.43
(0.8,1,2)	50	2.19	2.34	1.21	2.39	2.30	1.34	2.10	2.21	1.25
	100	3.42	2.96	1.83	3.64	2.97	1.86	3.81	3.09	1.67
	500	9.71	5.70	3.06	9.10	5.43	3.06	8.79	5.44	3.09
(0.2,1,3)	50	1.91	2.09	2.03	1.99	2.16	2.16	1.91	1.91	2.07
	100	2.93	2.92	2.41	2.57	2.65	2.22	3.07	2.88	2.57
	500	5.94	4.16	3.43	6.08	4.26	3.28	6.01	4.55	3.07
(0.5,1,3)	50	2.07	2.61	2.08	2.01	2.51	2.26	2.02	2.53	2.09
	100	3.02	3.20	2.71	3.08	3.11	2.68	2.97	3.02	2.64
	500	5.05	4.05	3.23	5.02	3.86	3.16	4.99	4.00	3.17
(0.8,1,3)	50	2.91	2.65	1.81	2.86	2.71	1.91	2.94	2.66	1.81
	100	4.55	3.27	2.43	4.53	3.23	2.31	4.51	3.21	2.51
	500	7.52	4.09	3.46	7.86	4.34	3.55	7.42	4.02	3.59
(0.2,1,5)	50	2.02	2.25	1.78	2.02	2.40	1.85	2.08	2.36	2.10
	100	2.37	2.44	2.01	2.50	2.71	2.14	2.40	2.46	1.92
	500	2.53	2.62	1.85	2.46	2.68	1.79	2.45	2.42	2.02

Table 3.18 (continued)

		α								
Parameter Vector		0.1			0.2			0.3		
	n	p_1	$ë_1$	$ë_2$	p_1	$ë_1$	$ë_2$	p_1	$ë_1$	$ë_2$
(0.5,1,5)	50	2.09	2.09	1.90	2.00	2.14	2.03	2.02	2.12	1.97
	100	2.18	2.22	1.96	2.10	2.17	2.04	2.13	2.29	1.98
	500	2.15	2.24	2.01	2.21	2.17	1.95	2.11	2.18	2.05
(0.8,1,5)	50	2.82	2.30	2.27	2.83	2.25	2.37	3.01	2.32	2.40
	100	2.97	2.39	2.37	2.90	2.26	2.48	2.88	2.29	2.59
	500	2.38	2.01	2.22	2.23	2.05	2.11	2.46	2.02	2.17
(0.2,1,8)	50	1.57	1.87	1.43	1.57	2.34	1.46	1.45	1.78	1.46
	100	1.50	1.71	1.51	1.47	1.70	1.45	1.41	1.78	1.40
	500	1.48	1.59	1.41	1.50	1.63	1.41	1.42	1.65	1.44
(0.5,1,8)	50	1.59	1.72	1.59	1.52	1.74	1.58	1.53	1.67	1.60
	100	1.54	1.68	1.66	1.59	1.73	1.58	1.51	1.66	1.68
	500	1.57	1.73	1.67	1.56	1.64	1.57	1.58	1.73	1.56
(0.8,1,8)	50	2.43	2.22	2.76	2.39	2.19	2.80	2.15	2.14	2.42
	100	1.99	2.14	2.50	2.03	2.05	2.40	2.02	2.20	2.38
	500	2.09	2.25	2.31	2.09	2.09	2.38	2.05	2.08	2.28
(0.2,1,10)	50	1.38	2.25	1.38	1.37	1.54	1.37	1.37	2.18	1.43
	100	1.34	1.56	1.29	1.38	1.59	1.37	1.33	1.65	1.30
	500	1.36	1.47	1.39	1.38	1.50	1.41	1.36	1.52	1.36
(0.5,1,10)	50	1.43	1.57	1.64	1.49	1.63	1.46	1.42	2.65	1.53
	100	1.50	1.58	1.53	1.50	1.55	1.53	1.56	1.54	2.62
	500	1.51	1.49	1.52	1.44	1.58	1.52	1.52	1.55	1.51
(0.8,1,10)	50	2.10	2.17	2.83	2.14	2.21	2.91	2.11	2.10	2.54
	100	2.13	2.12	2.53	2.09	2.10	2.66	2.10	2.11	2.43
	500	2.00	2.22	2.21	1.99	2.08	2.09	2.06	2.07	2.22

Table 3.19
Relative Mean Squared Errors. The entries are ratios of the mean squared errors of the parameter estimates when no supplementary information is available to those when the true subpopulation is known for a fraction á of the data

		α								
Parameter Vector		0.1			0.2			0.3		
	n	p_1	$ë_1$	$ë_2$	p_1	$ë_1$	$ë_2$	p_1	$ë_1$	$ë_2$
(0.2,1,2)	50	1.44	1.94	2.45	1.73	1.88	3.41	1.71	1.98	3.41
	100	2.37	3.38	3.53	3.82	3.63	5.29	2.92	3.69	3.73
	500	19.66	14.39	8.66	19.71	3.54	10.43	12.29	11.66	7.18
(0.5,1,2)	50	1.51	4.11	2.67	1.83	4.35	2.78	1.83	4.15	3.31
	100	4.47	6.84	4.49	4.23	7.18	3.59	4.20	7.45	3.94
	500	33.84	26.84	3.00	30.41	4.73	12.99	30.93	22.38	11.98
(0.8,1,2)	50	5.40	5.59	1.70	6.44	5.37	2.02	4.75	4.90	1.83
	100	13.03	8.96	3.92	15.00	9.08	3.77	16.55	9.83	3.35
	500	107.95	34.32	0.56	93.10	0.73	10.48	86.70	31.11	10.39
(0.2,1,3)	50	3.67	4.75	4.11	3.97	4.93	4.66	3.65	3.94	4.28
	100	8.62	8.70	5.85	6.59	7.05	4.91	9.47	8.47	6.60
	500	35.62	17.53	11.89	37.52	18.49	10.95	36.57	20.84	9.53
(0.5,1,3)	50	4.52	6.85	4.40	4.30	6.36	5.18	4.38	6.41	4.40
	100	9.82	10.62	7.42	10.02	10.09	7.20	9.35	9.46	6.96
	500	26.32	17.38	10.45	25.81	15.91	10.00	25.62	16.61	10.05
(0.8,1,3)	50	9.87	7.24	3.77	9.65	7.52	4.15	10.22	7.39	3.71
	100	24.87	11.33	6.49	24.70	11.18	6.06	23.83	11.09	6.64
	500	63.81	18.41	12.13	72.87	21.40	12.90	64.43	18.21	13.04
(0.2,1,5)	50	4.06	5.09	3.20	4.15	5.89	3.49	4.39	5.61	4.50
	100	5.62	5.94	4.05	6.26	7.34	4.57	5.80	6.05	3.72
	500	6.46	6.85	3.44	6.07	7.18	3.21	5.98	7.43	4.06

Table 3.19 (continued)

Parameter Vector	n	α = 0.1 p_1	0.1 $ë_1$	0.1 $ë_2$	α = 0.2 p_1	0.2 $ë_1$	0.2 $ë_2$	α = 0.3 p_1	0.3 $ë_1$	0.3 $ë_2$
	50	4.46	4.43	3.61	4.03	4.66	4.10	4.12	4.50	3.87
(0.5,1,5)	100	4.77	4.95	3.84	4.46	4.80	4.15	4.55	5.32	3.90
	500	4.61	5.04	4.06	4.88	4.74	3.81	4.46	4.75	4.19
	50	8.98	5.59	5.35	9.06	5.45	5.80	10.34	5.79	6.01
(0.8,1,5)	100	9.59	6.04	5.65	9.12	5.40	6.18	8.96	5.61	6.75
	500	5.85	4.12	4.90	5.14	4.28	4.48	6.19	4.15	4.71
	50	2.51	3.51	2.05	2.48	5.51	2.15	2.13	3.19	2.13
(0.2,1,8)	100	2.25	2.91	2.28	2.17	2.90	2.10	1.99	3.20	1.98
	500	2.19	2.53	1.99	2.25	2.66	1.97	2.02	2.73	2.06
	50	2.54	2.96	2.53	2.31	3.04	2.48	2.33	2.79	2.57
(0.5,1,8)	100	2.38	2.84	2.76	2.52	2.99	2.49	2.28	2.77	2.83
	500	2.45	2.98	2.79	2.45	2.72	2.45	2.50	3.01	2.43
	50	6.14	4.98	7.87	5.93	4.97	8.11	4.77	4.68	6.00
(0.8,1,8)	100	3.97	4.57	6.24	4.19	4.26	5.78	4.15	4.90	5.72
	500	4.37	5.07	5.33	4.38	4.38	5.64	4.20	4.33	5.20
	50	1.92	5.11	1.92	1.89	2.37	1.88	1.87	4.80	2.05
(0.2,1,10)	100	1.80	2.45	1.66	1.90	2.57	1.88	1.77	2.72	1.70
	500	1.87	2.16	1.94	1.91	2.24	2.00	1.84	2.32	1.85
	50	2.03	2.47	2.69	2.23	2.67	2.13	2.02	2.73	2.33
(0.5,1,10)	100	2.24	2.48	2.35	2.26	2.41	2.35	2.42	2.37	2.61
	500	2.27	2.23	2.30	2.07	2.50	2.31	2.31	2.40	2.28
	50	4.48	4.71	8.20	4.63	4.89	8.92	4.55	4.44	6.56
(0.8,1,10)	100	4.55	4.50	6.41	4.36	4.40	7.12	4.41	4.45	5.90
	500	4.01	4.94	4.90	3.97	4.34	4.36	4.24	4.30	4.95

The results of Tables 3.18-3.19 reveal the large gain achieved when some observations have known origin. When the proportion of known values is as small as 0.1, (which implies 5 observations in a sample of size n=50) the gain is large. In general, the gain in both efficiency and accuracy increases as the sample size and the proportion of known values increase.

However, we can see that this increase is not linear. Note also that the gain in accuracy is very large when the first component has a small mixing proportion.

Another interesting point is that the gain is greater for the estimation of the mixing proportion and the first component, especially when the components are not well separated.

In concluding this section we may note that the obtained results reveal an important fact. If additional information is available for a small proportion of the data, the improvement in the accuracy of the obtained estimates is very large. So, for example an insurance company can achieve larger gains if it obtains additional information for its clients however expensive this might be.

3.6.3 Example

Example 3.1 (continued) We will use some supplementary information concerning the dataset of Example 3.1 , describing the number of crimes in one month periods in Greece for every month from January 1982 to January 1994. Table 3.20 contains the original data, month by month.

Table 3.20
Number of crimes in one month periods in Greece (January 1982- January 1994)

	Year												
Month	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994
January	0	2	0	0	0	2	4	1	1	1	2	5	8
February	0	2	1	2	2	1	2	3	3	3	2	4	
March	3	1	1	4	1	1	1	3	4	5	2	4	
April	0	1	1	2	1	0	1	2	0	7	3	4	
May	0	0	0	1	1	2	4	3	0	4	5	5	
June	0	1	2	1	1	3	1	1	2	5	4	1	
July	1	2	2	1	1	0	3	1	2	6	2	5	
August	0	1	1	3	1	4	2	1	3	6	4	1	
September	1	0	2	2	0	2	0	1	4	1	9	2	
October	2	1	2	0	3	0	1	4	4	4	5	2	
November	2	2	4	2	1	2	3	0	4	6	5	3	
December	1	1	2	1	1	4	2	3	3	8	6	4	

Source: The Greek newspaper ‘TA NEA’ 15/2/1994

In section 3.3 a 2-finite mixture Poisson mixture was fitted to this data set. Trying to identify the two components, we can see that in the period 1989-1993 the numbers of crimes

are larger than the numbers of crimes in the period 1982-1988. In fact the cut-off point between these two periods is not clear, but it is reasonable to assume that the first year (Jan 1982-Dec 1982, 12 observations) belongs to the first component, while the last year (Jan 1993-Jan. 1994, 13 observations) belong to the second component. We suppose (for illustrative purposes) that these 25 values contain information about the mixing proportion (which is not surely true), i.e. suppose that we identified these observations and we found that 12 of them belong to the first component and the remaining 13 to the second component. We used this supplementary information to form an M2 type sample. From the total of 145 observations, 25 (almost 17%) belong to known subpopulations. Applying the EM algorithm described above we obtained the estimates $\hat{p}_1$ =0.5119 (0.084), $\hat{\lambda}_1$=1.2026 (0.135), $\hat{\lambda}_2$=3.3309 (0.328), where the numbers in the parentheses represent the jacknife standard errors of the parameters. We can see that these estimates differ from the estimates when no additional information was available.

We will not pursue further this example, mainly because it is artificial. However an interesting problem related to this data set is to examine if there is a point which partitions the whole period into two subperiods.

3.7 Semiparametric Maximum Likelihood Method for Mixtures

3.7.1 Introduction

So far, we have been concerned with the case of k-finite mixtures where the number of components is known a priori. In practice the case of semiparametric ML method is of special interest. As semiparametric we call the case where the number of support points is not known a priori and it must be estimated from the data as well. In this case we maximise the likelihood over all the mixing distributions with finite support.

The importance of such methods is vital because:

- As Laird (1978) has shown, if the true mixing distribution is continuous we are restricted to estimate the mixing distribution by a finite-step distribution, i.e. by reducing the mixture model to a finite mixture model with unknown number of support points.
- The number of support points itself is of special interest in many applications as it determines the number of subpopulations comprising the entire population.

Clearly, this case is far more complicated than the case of known k and special algorithms and numerical methods are needed. We will give a fundamental theorem proved by Lindsay (1983a) as well as related results which can give us information about the number of support points and the conditions which ought to be satisfied from the semiparametric ML estimates of the mixing distribution. Several algorithms proposed for obtaining the ML estimates will be discussed critically. The case of Poisson mixtures will be treated in depth, since the available algorithms are known to have some impediments in practical situations.

3.7.2 Conditions for the Existence of the Maximum Likelihood Estimate

Lindsay (1983a,b) described the case of semiparametric ML (hereafter SML) estimation for finite mixtures. He gave the general theorem for ML estimation which is the basis for many of the methods which we will describe in the sequel.

Again, the gradient function defined in (3.6), plays an important role. The applicability of the theorem lies in the fact that it gives sufficient conditions for examining if the global maximum has been obtained. Such conditions will be extracted in more comprehensive forms, later in chapter 6.

Lindsay 's (1983a) theorem is as follows:

Theorem 3.3 (Lindsay, 1983a). $\hat{G}$ is the SML estimate of the mixing distribution G iff the following relations hold:

a) $D(\theta, \hat{G}) = 0$ for each è which is a support point of $\hat{G}$

b) $D(\theta, \hat{G}) \leq 0$ for all other values of è, not in the support of $\hat{G}$.

This theorem extends the results of Whittle (1973) for D-optimal designs to the case of mixture models. The proof of the theorem has been given in Whittle (1973) for a more general case. The importance of Theorem 3.3 for mixtures lies on the fact that it provides us with sufficient conditions for an estimate to be an SML estimate. Note that Theorem 3.3 applies also when the vector of parameters *è* is multidimensional, as for example in the case of finite normal mixtures with unequal variances.

From this we can see that all the points in the support of the SML estimate are the local maxima of the gradient function. Comparing Theorem 3.1 with Theorem 3.3 we can see that the difference is that in the case of restricted support size the support points can also be

minima or saddle points of the gradient function while for the case of SML estimation they ought to be maxima.

Theorem 3.3 provides useful tools for calculating the SML estimate. A natural approach is to apply the EM algorithm for successive values of k. For each value of k we check if the conditions are satisfied, otherwise we proceed with the next value of k. It is interesting that the likelihood can be maximised with few support points and the addition of a further support point will not increase the likelihood. This is the case when the likelihood function for fixed k, has multiple maxima, making the ML estimate inconsistent. Pfanzagl (1988) discussed the consistency of the ML estimates for mixture models. Before examining more thoroughly the conditions we discuss another important issue: the uniqueness of the SML estimator.

The uniqueness of the SML estimator has been showed by Simar (1976) for the case of Poisson mixtures. Lindsay (1983a,b) and Lindsay and Roeder (1992,1993) showed that the SML estimator is unique for members of the continuous exponential family. They also showed that the SML estimator is unique for discrete mixtures if and only if the probability distribution evaluated using this SML estimator of the mixing distribution does not coincide with the observed relative frequency distribution. Obviously for discrete distributions with support in the positive axis, like the Poisson distribution, the probability function evaluated using the SML estimate of the mixing distribution will give positive probability to values greater than the maximum observed value. Hence this estimated probability distribution will not coincide with the observed relative frequency distribution.

Let us return to the conditions of Theorem 3. Bohning *et al.* (1994) proposed to check for the conditions of Theorem 3 by choosing a large number of values in a reasonable interval and checking if the maximum of the gradient function is a value very close to 0 that occurs on the support points of the SML estimate. This because small perturbations due to the computer accuracy may not allow the researcher to calculate a value which is exactly 0. It is clear that such an approach, for checking if the SML estimate has been found, is time demanding and simpler conditions are needed. If the maximum has been obtained adding one new support point will not increase the likelihood and hence the likelihood ratio test statistic will take a zero value.

The results of Theorems 3.1 and 3.3 provide useful guides for checking if the global maximum is obtained. If we plot the gradient function, this ought to have 0 values at all points on the support of the solution, and if the global maximum is attained, it ought to be restricted

down the 0 line. Hence, by plotting the gradient function we can check if the solution is the SML solution. Otherwise, if there exist points outside the support of the solution, for which the gradient is 0, or points with a positive gradient, we have to add points because the global maximum has not been attained.

3.7.3 The Number of Support Points

From the above discussion it is evident that the ML estimate for a k-finite mixture is not necessarily the SML estimate. It is just the best possible solution with the given number of support points. The natural question at this point is whether we know something about the number of support points.

The answer is in the affirmative. Simar (1976) was the only one who concentrated on the particular case of ML estimation for Poisson mixtures. He provided the following theorem concerning the number k of support points.

Theorem 3.4 (Simar, 1976) . If k denotes the number of support points of the SML estimate of the mixing distribution, and N represents the largest observed value then:

a) If $\lambda_1 = 0$ then $k \le \left[\frac{N+2}{2}\right]$, while if $\lambda_1 > 0$ then $k \le \left[\frac{N+1}{2}\right]$.

b) In every case $k \le q$

where [a] is the integer part of a, and q is the number of distinct values in the sample .

Laird (1978) conjectured that the number of support points in mixtures from continuous densities cannot be larger than the sample size. She also gave an interesting guide to this search. For a mixed distribution the problem of counting the number of support points is equivalent to counting the number of modes of a mixture of n conjugate densities. For example, for the case of the Poisson probability function, the Gamma density is the conjugate. So if we have assumed a mixed Poisson probability function, then we take a mixture of n Gamma distributions, with parameters $x_i + 1$ and 1 respectively, $i=1, \ldots, n$, and we count the number of modes. This approach gives us an upper bound for the number of support points.

Lindsay (1983a) proved the conjecture of Laird (1978), namely that the number of components cannot be larger than the sample size. Lindsay and Roeder(1993) gave a result similar to Simar's for general discrete distributions.

Intuitively, when we try to estimate a distribution with k support points the number of estimated parameters is 2k-1. If we have observed only N classes (different values), then with N parameters we can theoretically fully reconstruct the observations (because we simply solve a non-linear system of N equations with N unknowns). So we need to restrict the number of support points. Adding one more component implies that the unknown values to be estimated are more than the estimating equations and this leads to intractabilities. Note that since we want to maximise the likelihood and not to solve the system of equations explicitly, the problem lies in that the constraints for the maximisation are too many.

The above results on the number of support points are a useful guide when searching for the SML estimate. On the other hand, when the mixture is discrete, this restriction prevents us from estimating the continuous mixing distribution with a finite approximation with many support points and hence closer to a continuous one. A simple example is the case of a Gamma mixing distribution for a mixture of the Poisson distribution. This leads to the negative binomial distribution. If the mean is not large, the estimation will provide us with an estimate with a few support points which will not resemble the true Gamma mixing distribution.

We will now describe algorithms designed for finding the SML estimate of the mixing distribution.

3.7.4 Algorithms for Semiparametric Maximum Likelihood Estimation for Mixtures

Special algorithms are needed for obtaining the SML estimate of the mixing distribution. The number of support points is unknown and this complicates the procedure. A simple answer to this problem is to derive the ML estimate for successive values of k. This can be easily carried out via the EM algorithm of the previous section. Conditions of Theorem 3.3, can be used as a stopping criterion to this approach. This simple algorithm is very useful in practice but it may require a lot of computational effort since, as we saw, the EM algorithms require a lot of computational work at each step (for each value of k). Jewell (1982) proposed this method with continuous use of the EM algorithm for mixtures of the exponential distribution. The case of mixtures of the Poisson distribution can be handled in a similar way. Theorem 3.4 provides an upper limit for the number of components.

More sophisticated algorithms have been proposed in the literature, using special methods of numerical analysis. The main idea for them is to start with an initial solution with

a few support points, usually one or two, and then add one or more new points at each step, sometimes replacing old but "bad" points till some criteria are fulfilled.

In the sequel, we describe the algorithms for the mixed Poisson case.

3.7.4.1 The Vertex Direction Method (VDM)

The idea is to start with some initial value, say P^i which, in general, represents the estimate of the mixing distribution after i-steps, and then add as a new point the value of *è* which maximises the gradient function given in (3.6). The probability associated with this new support point must be calculated in such a way so that for the new estimate P^{i+1} the loglikelihood is better, namely $\ell(P^{i+1}) \geq \ell(P^i)$. Generally $P^{i+1} = (1-a)P^i + aP_\theta$ where P_θ is a distribution which puts all its mass at the point è, i.e. a degenerate distribution. Clearly, *á* is the probability assigned at the new support point.

So the VDM algorithm consists of the following steps.

Step 1 : Find $\theta_{\max}$ to maximise $D(\theta, P^i)$

At this point we want to find a new "good" point of support. The quantity $D(\theta, P^i)$ is the directional derivative. So, by maximising $D(\theta, P^i)$ we find the best point to the direction towards the new estimate P^{i+1} ; $\theta_{\max}$ is the new support point.

Step 2 : Find á to ensure that $\ell\left((1-a)P^i + aP_{\theta_{\max}}\right) \geq \ell(P^i)$. At this step we construct the new estimate, adjusting the probabilities of the "old" support points so that the new probability estimates add up to 1.

Step 3 : Examine if a global maximum is attained using the conditions of Theorem 3.3, otherwise go back to step 1.

Obviously this method requires a lot of numerical work at both steps 1 and 2. At step 1, maximisation can be achieved by initially searching a grid of distinct points in some interval. A good choice for such an interval is between 0 and the maximum observed value in the sample. Then one may start from the point where the gradient function has its maximum to locate the maximum by some iterative scheme like the Newton-Raphson. The strategy is that

the grid search reaches the maximum which is then located easily via a standard maximisation algorithm. The process is carried out by searching for the point where the derivative is 0. However, the algorithm may lead to the minimum near the point instead of the maximum itself. So, we have to check if the obtained point is a maximum and not a minimum.

The main problem associated with this step, is that the maximum may lie outside the admissible range. It may be negative or the gradient function may increase to the infinity. In both circumstances the maximum cannot be found and the algorithm stops.

In step 2 we need to find the value of á. Bohning (1989, 1995) describes algorithms for finding a value for á. He calls these algorithms as monotone step algorithms. He shows that the problem of finding the value of á, can be reduced to a problem of estimating a closed area. So algorithms used for estimating an area are useful for finding a value for á.

Apart from the monotone step algorithms another choice would be to find an á which maximises $\ell(P^{i+1})$ with respect to á. Bohning (1995) shows that $\ell(P^{i+1})$ is concave with respect to á and thus a maximum value exists which is very easy to locate by a numerical algorithm. Note that the formulas which stem from the monotone step-length algorithms are in fact the first step for iterative numerical methods for solving an equation with initial value equal to 0. For example if $\phi(a) = \ell(P^{i+1})$ then the Newton-Raphson proposal for a monotone step length algorithm is to choose $a = -\phi'(0)/\phi''(0)$ where $\phi'(0)$ and $\phi''(0)$ are the first and second derivatives of $\phi(a)$. Since the maximisation of $\phi(a)$ is equivalent to solving for $\phi'(0) = 0$ the above formula is the first iteration for a Newton-Raphson method with initial value $a = 0$.

Problems may occur if the value á which maximises $\phi(a)$ lies outside the interval [0,1]. Bohning (1995) suggested the use of different monotone step-length algorithms for cases when this failure occurs.

In case we have a restriction for the number of support points we must add this condition at step 3. This is true for the Poisson case.

Note that Lindsay (1983a) gave bounds for the improvement of the likelihood between successive distributions with k and k+1 points. He also proposed the use of these bounds as terminating conditions for the iterations.

Simar's algorithm is a modification of the VME. Lesperance and Kaldbfeish (1992) call this algorithm Modified Vertex Direction method (MVME). The difference is that at step

2, instead of adjusting the probabilities for the old points and estimating the probability for the new support point, he proposed estimating the probabilities for all the points maximising the likelihood. The EM algorithm is plausible to provide the estimates. Simar proposed some ways to help the search for the maximum. In case where the addition of a new support point leads to a non admissible number of support points, he proposed to find an admissible solution by solving a moment problem with a number of support points in the admissible range and then continue the iterations. Simar did not prove the convergence of his algorithm. Bohning (1982) showed that under mild conditions the method really converges to the global maximum, and he suggested some modifications of the algorithm to improve the convergence.

The algorithm itself has some serious disadvantages.

The first is that it is very slow. It converges to the maximum with a very big effort. Bohning (1995) proposed some improvements for the VDM. These improvements however, had a marginal effect because of some inherent disadvantages.

As can be seen for this algorithm, the initial value (or values) are of great importance. "Bad" initial values can destroy the algorithm. This is due to the fact that at step 1 we are not able to find a maximum because the quantity $D(\theta, P^i)$ is monotonic or the maximum is outside the admissible range. On the other hand, the initial point remains at the estimated mixing distribution forever, because we just add points. This may cause the destruction of the algorithm. For the Poisson case where the number of support points is usually small, the algorithm is not satisfactory since we have to estimate very few support points. (see, for example, Brannas and Rosenqvist, 1994). For other cases, i.e. mixtures of normal or other continuous distributions, where there is no such a strict limitation for the number of support points, the initial value almost disappears because we add a large number of new points and the effect of the initial point is negligible.

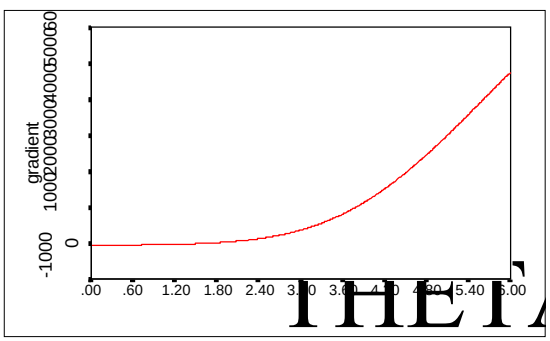
Let us now describe another problem stemming from the results of section (3.2). It is evident from Theorem 3.3 that the mean of the estimated mixing distribution ought to satisfy the first moment equation, i.e. the mean of the estimated mixing distribution should be equal to the one estimated from the sample. This is due to the fact that the gradient function and its derivative evaluated at the support points ought to be 0, as Theorem 3.1 shows. This complicates the steps of the VDM. To see this we will treat the general case of mixtures from the exponential family. Recall that $\mu(\theta)$ represents the mean value reparameterization. Suppose that at this moment we have k points, say $\mu(\theta_j)$ with associated probabilities p_j , for

$j=1, \ldots, k$. Then, $\sum_{j=1}^{k} p_j \mu(\theta_j) = \bar{x}$ where $\bar{x}$ is the sample mean. Thus the new support point, say $\grave{\imath}(\grave{e}_{k+1})$, will be assigned a probability á whose value is such that the increase in loglikelihood between the two models with k and k+1 points is maximised. If the new point is $\grave{\imath}(\grave{e}_{k+1})$, the condition $(1-a)\sum_{j=1}^{k} p_j \mu(\theta_j) + a\mu(\theta_{k+1}) = \bar{x}$ ought to be satisfied, as this equation is one of the estimating equations for the ML estimation with k+1 support points. In any other case the increase in the loglikelihood will not have been maximised. Solving with respect to á we obtain á=0, which implies that we reject the new support point. Any other choice of á would lead to a solution which is not an ML solution with k+1 support points.

This reveals the following problem: the new point will always de dropped if the solution satisfies the mean equation. On the other hand, if we avoid such solutions, i.e. if our solutions do not satisfy the mean equation, then the resulting estimate is not a ML estimate.

Example 3.1 (continued). Suppose that we start with the one support point at *ë*. Figures 3.3 depict the gradient function for two different choices of *ë*. Note that if $\lambda < \bar{x} = 2.24$, the gradient function tends to infinity, while if $\lambda \geq \bar{x} = 2.24$, the gradient function has maximum at 0. The shape for other values of ë is quite similar if *ë<2.24,* i.e. if *ë* is smaller than the sample mean, the gradient function tends to infinity, otherwise it has a maximum at 0. The results discourage the use of the VDM method since the maximum does not exist or it is equal to 0, and thus we cannot select a second point. In fact, in the second case the maximum occurs in the negative axis, which is not acceptable because the parameter of the Poisson distribution must be positive. This demonstrates the problematic behavior of the gradient function when starting with one support point.

a



b

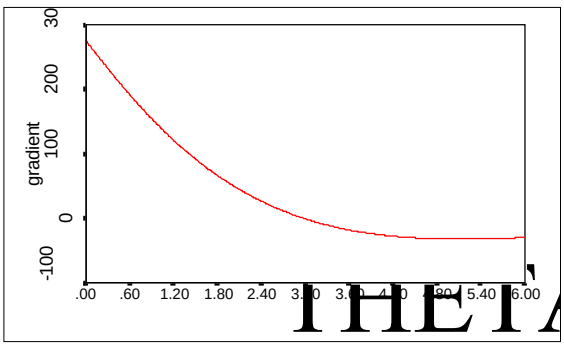


Figure 3.3 The gradient function when starting with one support point at a) ë=1 and b) ë=3 respectively.

Suppose now, that we start from a 2 point solution in order to overcome this difficulty. Figure (3.4) shows that the behavior of the gradient function is now much better. We used 7 different cases and the maximum exist for all the cases. These 7 cases are presented in Table 3.21 with the new support point. For the case (g), which is the ML estimate for this dataset, no other point can be added because we have already obtained the SML estimate. The support points are the two local maxima of the gradient function and for no other point is the gradient function equal to 0.

From Table 3.21 we can see how different are the new estimates when a new point is added.

Table 3.21
The values of the support points used for calculating the gradient functions of figure 3.4, with the new support point, i.e. the maximum of the gradient function

	initial estimate				support points after the 1st iteration		
case	p_1	$ë_1$	$ë_2$	new point			
a	0.25	1	4	0.72	0.72	1	4
b	0.5	1	4	1.74	1	1.74	4
c	0.8	1	7	3.64	1	3.64	7
d	0.672	1.8	6.7	0.06	0.06	1.8	6.7
e	0.5	1	3	7.42	1	3	7.42
f	0.9	2	3	8.16	2	3	8.16
g (MLE)	0.672	1.488	3.788	-			

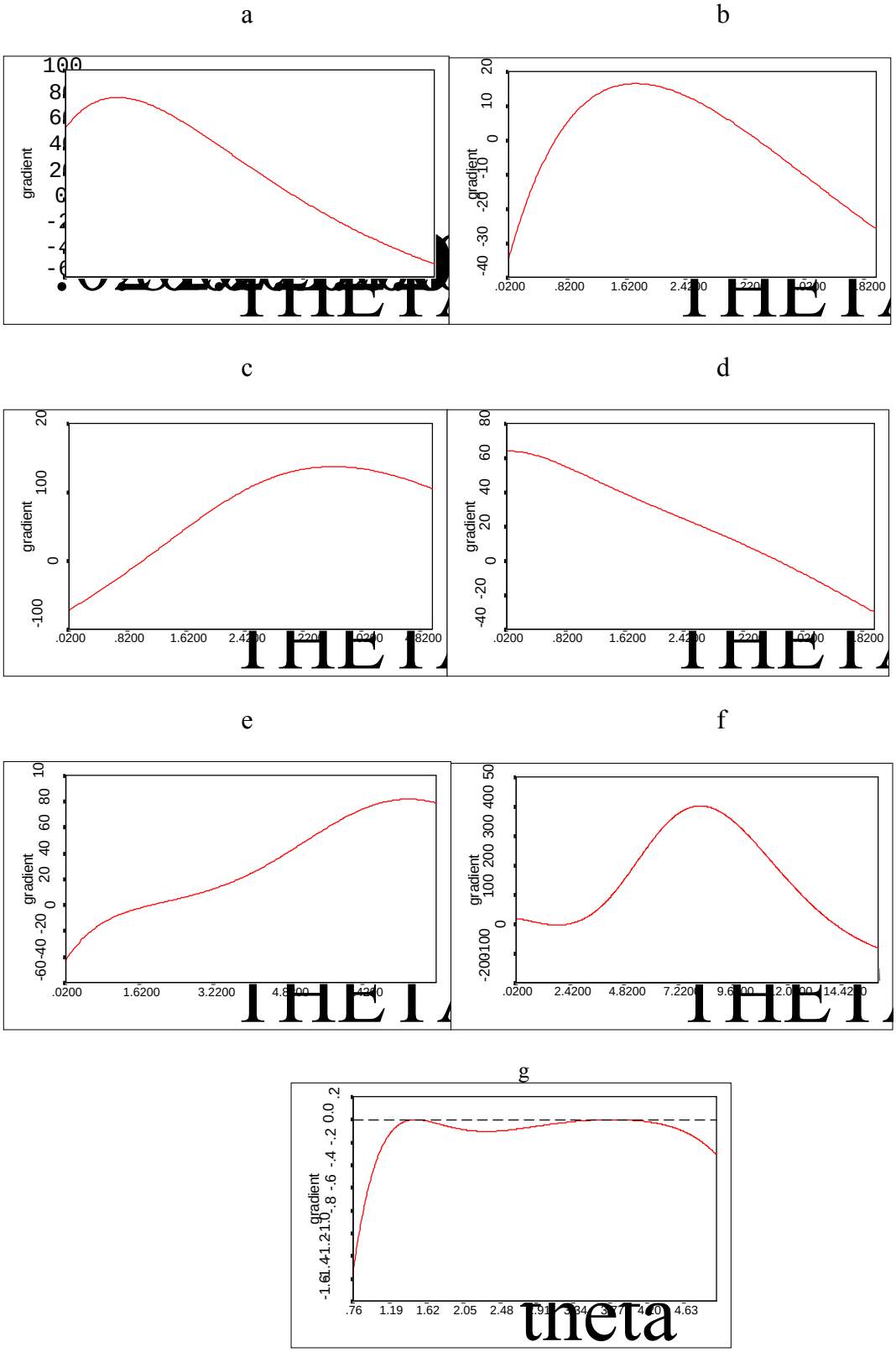


Figure 3.4 Plots of the gradient function when we try to add a third point. The 7 2-finite Poisson mixtures used had parameter vectors: a) (0.25,1,4), b) (0.5,1,4), c) (0.8,1,7), d) (0.67,1.8,6.7), e) (0.5,1,3), f) (0.9,2, 3) and g) (0.672, 1.488, 3.788) which is the ML estimate.

However some difficulties are again present. We run the VDM algorithm for the data of Example 3.1, starting from a 2-point solution with parameters p_1=0.5, $\ddot{e}_1$=1.12 and $\ddot{e}_2$=3.38. The results, after 48 iterations, gave the 50-point solution presented in Table 3.22. The loglikelihood for the solution with 50 support points, (ignoring the restriction on the support points) was -274.242 which is not the maximised loglikelihood. Figure 3.5 provides the histogram of this estimated mixing distribution. The majority of the components are close together. The results clearly discourage the use of the standard VDM algorithm for Poisson mixtures. For continuous mixtures, we hope that adding support points to the estimate of the mixing distribution will lead to a smooth estimate of the mixing distribution. Moreover, in this case we can add many support points without any problem.

Table 3.22
The estimated mixing distribution of the data of example 3.1 with 50 support points

ë	p	ë	p	ë	p	ë	p
1.128	0.21329	1.660	0.01135	3.880	0.00388	4.280	0.00355
1.510	0.00834	1.680	0.07222	3.890	0.00391	4.390	0.00345
1.520	0.03437	1.690	0.01203	3.900	0.00389	4.530	0.00340
1.530	0.02646	1.740	0.01335	3.920	0.00388	4.580	0.00105
1.540	0.01806	1.750	0.04298	3.940	0.00388	4.720	0.00347
1.550	0.00921	1.770	0.02323	3.950	0.00387	4.820	0.00336
1.560	0.01893	1.830	0.01810	3.970	0.00388	5.070	0.00336
1.570	0.03002	2.200	0.07504	4.000	0.00385	5.380	0.00334
1.580	0.00980	3.385	0.21329	4.040	0.00378	5.460	0.00157
1.590	0.00999	3.840	0.00780	4.080	0.00378	5.650	0.00309
1.610	0.01043	3.850	0.00778	4.130	0.00516	6.080	0.00337
1.620	0.01055	3.860	0.00393	4.200	0.00363	6.540	0.00418
1.640	0.01097	3.870	0.00390				

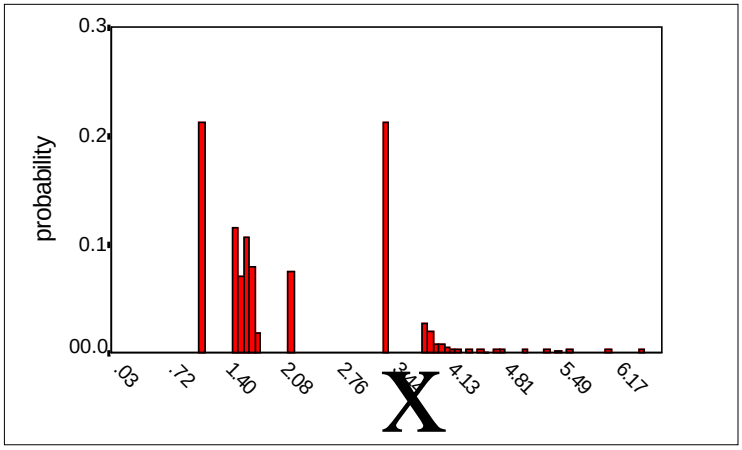


Figure 3.5 The estimated mixing distribution via the VDM algorithm for the data of example 3.1.

A way of avoiding such difficulties may be the use of the VDM in connection with the EM. The algorithm can be described as

Step 1 : Find the new P^{i+1} with the VDM

Step 2 : Make some iterations with the EM algorithm, using P^{i+1} as initial values. Go back to step 1.

With this combination the problem with the initial values is hoped to disappear. The convergence is again guaranteed, because of the monotone nature of the EM algorithm. The contradiction is that since the EM algorithm is used, no substantial gain is achieved compared to the simple method of successive EM methods. Recall the difficulties in assigning a probability to a new support point if the solution is ML and hence the moment equation is satisfied.

Concluding we can say that the applicability of the VDM algorithm is problematic, since it cannot get rid of the initial points and when the number of support points is restricted, it adds redundant points.

3.7.4.2 The Vertex Exchange Method (VEM)

The Vertex Exchange Method (VEM) tries to overcome some of the disadvantages of the VDM. At each step it adds a new support point but it detects if there is a "bad" point which is extracted. So step 1 is followed by a step where we find the "worst" support point from the "old" support points and examine if the new point $\theta_{\max}$ can replace at all the "bad" point $\theta_{\min}$. In this case the "bad" point is eliminated. Lesperance and Kaldbfeish (1992) described the algorithm in detail. The algorithm consists of the following steps.

Step 1 : Find $\theta_{\max}$ to maximise $D(\theta, P^i)$ over all possible values of è. This step leads to the new support point. Grid search with a complementary numerical search is a useful tool for finding it.

Step 2 : For all the points in the support of P^i calculate the gradient function $D(\theta, P^i)$. Find the point $\theta_{\min}$ which has the minimum value over all the support points.

Step 3 : Set $P^{i+1} = P^i + \alpha p^* \left(P_{\theta_{\max}} - P_{\theta_{\min}} \right)$, where p^* is the probability of the "bad" support point. The meaning of this expression is that we take some proportion á of the probability of the "bad" support point and we assign it to the new support point. If á=1 , we reject the bad support point. If á=0 we do not change our estimate at all. The problem is again

to find the value of á to ensure that $\ell(P^{i+1}) \geq \ell(P^i)$. Again, a monotone step-length algorithm or direct maximisation are possible methods.

Step 4 : Examine if the maximum is obtained.

With the VEM method it is not necessary to add a point at every iteration. However, if the new point is "bad" then the algorithm will fail. This method is more dynamic than the VDM and it converges quicker than the VDM. Lesperance and Kalbfleisch gave examples to show the superiority of the VEM algorithm. Bohning (1995) suggested some slight improvements of the algorithm.

A problem which may occur with the VEM method relates to the choice of initial values. "Bad" initial values may lead to "bad" choices of new points and since the algorithm exchanges one point with another at each iteration, it may delay very much to get rid of the bad points. For the Poisson case, where the number of admissible support points is small, the VEM algorithm does not give satisfactory results.

Example 3.1 (continued) We run the VEM algorithm for the data of Example 1. A solution with 2 support points was used again as an initial estimate. Specifically, the initial values were set to p_1=0.5, $ë_1$=1.12 and $ë_2$=3.38 (the same with the initial values used for illustrating the VEM algorithm).

Table 3.23 contains the estimated mixing distribution with 64 support points, derived after 100 iterations. Figure 3.6 depicts the estimated mixing distribution. Again the loglikelihood is -274.241 which is inferior to the ML solution with 2 support points (which is the SML).

Table 3.23
The estimated mixing distribution with 62 support points derived using the VEM algorithm for the data of the example 3.1

ë	p	ë	p	ë	p	ë	p	ë	p
1.128	0.26645	1.585	0.00169	1.720	0.01166	2.030	0.05401	5.025	0.00202
1.395	0.01034	1.590	0.00253	1.725	0.00068	3.385	0.31957	5.060	0.00146
1.425	0.01522	1.595	0.00407	1.735	0.00127	4.865	0.00108	5.080	0.00116
1.455	0.02160	1.600	0.00185	1.750	0.02821	4.915	0.00110	5.110	0.00116
1.515	0.00344	1.610	0.00068	1.785	0.00095	4.920	0.00108	5.135	0.00479
1.545	0.00355	1.620	0.00112	1.810	0.01628	4.925	0.00085	5.235	0.00166
1.550	0.02999	1.625	0.00120	1.815	0.03357	4.935	0.00161	5.255	0.00132
1.555	0.00197	1.645	0.00795	1.825	0.00161	4.940	0.00112	5.280	0.00131
1.560	0.00435	1.655	0.00265	1.860	0.00181	4.955	0.00114	5.300	0.00359
1.565	0.00151	1.665	0.00159	1.880	0.00063	4.970	0.00303	5.315	0.00207
1.570	0.00673	1.670	0.00818	1.890	0.04233	4.975	0.00082	5.345	0.00134
1.575	0.00088	1.675	0.03619	1.910	0.00212	4.990	0.00195	5.355	0.00023
1.580	0.00502	1.695	0.00133	1.965	0.00211	5.005	0.00386		

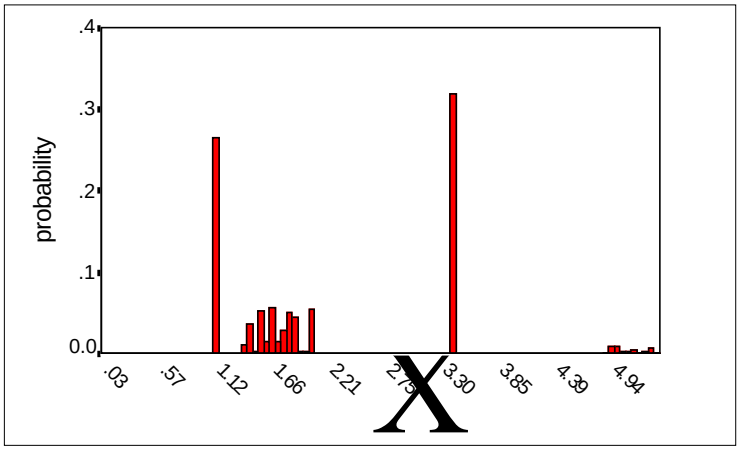


Figure 3.6. The estimated mixing distribution via the VEM algorithm for the data of example 3.1.

Another natural way of improving the results would be to use the algorithm in combination with the EM algorithm. After each VEM iteration some iteration of the EM algorithm may help the search. The comments of the previous section apply here as well. With such an approach the new components are usually dropped due to the results of section (3.2.2).

3.7.4.3 The Intra Simplex Direction Method (ISDM)

The VDM and VEM algorithms and their modifications have the computational disadvantage that one must keep track of and perform computations over the complete

accumulated set of support points at each iteration. All of them add at the most one new support point.

Lesperance and Kaldbfeish (1992) proposed another method. By their method at each step we find several new points instead of one. At step 1, instead of finding the global maximum, we find several local maxima points $\theta_1^*, \theta_2^*, \ldots, \theta_r^*$. Then we must find the probability assigned to each point $\theta_1^*, \theta_2^*, \ldots, \theta_r^*$ maximising the corresponding likelihood as in step 2 of the VDM and VEM algorithms. This method requires a lot of computational work. For example, we have seen in Figure (3.4) that the gradient function can have more than one local maxima. In general, it is difficult to obtain all the local maxima and we need a very careful search to do so. In fact, the added labour is in step 2, because from the grid search for the maximum of step 1 we have already calculated the gradient function for several values of è. The EM algorithm is an adequate choice for finding the probabilities of step 2. As Lesperance and Kaldbfeish (1992) point out the complicated computations at each iteration are compensated by the smaller number of iterations until the global maximum is attained. Again, this approach is not appropriate for the Poisson case since, as we have seen, the number of iterations is usually small. This method is known as the Intra-Simplex Direction Method (ISDM).

3.7.4.4 Related algorithms

Dersimonian (1986,1990) proposed an algorithm similar to Simar's that uses the conditions given by Lindsay (1983a) for examining if the maximum is obtained. Her algorithm treated the case of mixtures of normal, exponential, binomial and Poisson distributions, starting from a uniform estimate. Shee assigned equal probabilities to equally spaced points. It uses the EM algorithm until some kind of convergence is achieved and then, by maximising the gradient function, it finds a new support point. Then the EM algorithm is applied so as to maximise the likelihood for the new set of support points. The algorithm stops when we cannot add a new support point or the conditions of Lindsay (1983a) are satisfied.

An interesting connection with algorithms used in the field of D-optimal designs is discussed by Bohning (1989,1995). He showed that searching for the ML estimate of a mixing distribution is equivalent to searching for a D-optimal design. So results from this field are applicable. This is the reason why, in some cases, the algorithms appear with different names. The author cited a large number of references. In the same paper he described in detail the

monotone step-length algorithms. Lindsay (1983a,b) also described the similarity with D-optimal designs.

Mallet (1986) used a very similar approach to estimate the distribution of the random coefficient of a regression model, inducing his problem to a problem of estimating the mixing distribution. He also gave an interesting application. Some monotone step-length algorithms are also provided.

Heckman and Singer (1984) used the SML estimate for estimating the mixing distribution in duration models. They showed how much sensitive is the estimation based on a specified mixing distribution and they proposed the SML estimate as a method to avoid biasing the results by choosing an arbitrary mixing distribution.

Bohning (1989) described the geometry of the likelihood of mixtures. Similar is the work of Lindsay (1983a,b). They both showed pictorially, in a few dimensions, how we proceed to maximise the likelihood, giving an excellent insight to the whole procedure. Also this approach gives some knowledge about how we can improve our search. Bohning (1989) proposed some methods to do so, but we will not examine them in detail. Lindsay (1983a,b) connects the problem of ML estimation to some geometrical concepts and using known results from geometry he proves many useful properties. Much of related material can be found in Lindsay (1995). Certain properties of the ML estimator were shown to be geometric properties of the likelihood set. In the paper the author also gave some bounds for the possible improvement of the loglikelihood in each step of the VDM algorithm.

Constrained maximisation is described by Lesperance and Kalbfleisch (1992), Bohning (1995) and Susko *et al.* (1997). Lesperance and Kalbfleisch (1992) proposed a semi-infinite programming routine for the maximisation which seems to work well.

3.8 Properties of the Semiparametric Maximum Likelihood Estimate of the Mixing Distribution

In the case of fixed k, asymptotic variances and covariances can be computed using standard asymptotic theory. Unfortunately, the computation of the variances in the case of the SML estimation is not straightforward. The efficient scores needed for calculating asymptotic variances, though easily calculated in the case of the simple ML estimation, cannot be obtained easily for the SML estimator. Since now the number of parameters is not known, the exact calculation is not possible (Lindsay and Roeder, 1995). However, there is a case in latent models where this is possible, based on the connection of the resulting mixture model to

the conditional ML in the homogeneity model (see Lindsay *et al.*, 1991). A reasonable approach proposed by Lambert and Tierney (1984) is to estimate the standard errors from the formulas of the ML estimator when the number of support points is fixed, keeping in mind that the bias is asymptotically negligible.

The SML estimator possesses some interesting properties.

Simar (1976) showed the consistency of the SML estimator. Tierney and Lambert (1984) showed that functionals of the mixed distribution can be estimated consistently by functionals of the ML estimator of the mixed distribution, and these functionals are asymptotically normally distributed. The authors showed that the commonly used functionals of the empirical density are more efficient than the functionals based on the SML estimator. However, the difference is very small, so the authors advocate the use of functionals based on the SML estimator because issues other than the asymptotic efficiency are more interesting in some cases. In the second part of their article they examine the behaviour of the SML estimator in the case of a mixed Poisson distribution. Later, Pfanzagl (1988) treated in greater detail the consistency of the SML estimators for the mixture model case, while recently Van de Geer (1995) and Van Der Vaart (1996) discussed the asymptotic normality of the ML estimators for mixture models and for functionals related to it.

Lambert and Tierney (1984) showed the following properties:

Let q_i be the estimated frequencies for the observed data using the SML estimate and p_i the observed frequencies. The following results hold:

1. q_i is closer than p_i to the true values of the population
2. p_i and q_i have equal asymptotic variances in the nonboundary cases
3. q_i is smoother than p_i (something very important when we are interested in the tail of the distribution as for example, in actuarial applications).
4. The quantity $n^{1/2}\left(q_i - f_i\right)$ is asymptotically normally distributed, where n is the sample size and f_i is the true value of the population.

The authors judged that, despite the smaller standard error of the empirical frequencies, the ML estimate possesses some other appealing properties and thus its use is recommended.

A very useful result is that the standard error of q_i can be computed asymptotically as $q_i\left(1-q_i\right)/n$. This simple formula enables us to find confidence intervals for our estimates, or better confidence regions for the estimated mixed distribution.

Note however, that all the results of Lambert and Tierney are based on the assumption that the number of support points is unknown prior to the estimation and that the number of support points is large enough. In this case, when the estimated mixing distribution is supported by a few points, the asymptotic standard errors can be computed assuming that the number of support points is fixed. This approach should underestimate the standard errors. As the above authors showed, the bias is very little.

Harris (1991) showed that the estimated frequency of zero is always greater than the observed frequency and thus it is biased. This result can be used to check if the ML estimate has been obtained.

The results of section 3.2.2 reveal that the mean of the estimated SML estimate of the mixing distribution is the same as the sample mean. To see this, recall that we showed that the mean of the estimated mixing distribution, when k is fixed, is necessarily equal to the sample mean. This proof was based on the likelihood equations, which are equivalent to the conditions that the gradient function is 0 at the support points and has zero derivatives at these points. These conditions hold for the SML estimate, too. Hence, the mean of the estimated SML estimate equals the sample mean.

3.9 Conclusions

A natural question that arises after the description of all the algorithms is the appropriateness of each algorithm as well as a comparative judgement of the algorithms. It is clear that there is not a simple answer to this problem.

For fixed support size, the EM algorithm, with some modifications to improve its performance, is the best solution. Contemporary computer devices can facilitate the use of the EM algorithm. Despite its slow convergence, the time required is almost negligible. But what can we say about the case of flexible number of support points?

The question can be formulated in a different way. Do we want to restrict our attention to the admissible case, namely that k must be almost half of the distinct values in our sample? If the answer to this question is yes we have a lot of problems. We have to try a lot of different initial values, or even algorithms, to find the maximum and a careful examination is needed in order to be sure that the SML estimate has been obtained.

Another relevant problem arises when we have assumed a continuous mixing distribution: Is it relevant to try to estimate it by a finite step mixing distribution with so few

support points? Several authors conclude that the number of support points is evidence about the number of components. This is true for finite mixtures, but if the mixing is continuous the conclusion is misleading.

Suppose that we generate data from the negative binomial distribution. Then we estimate the mixing distribution, which in fact is a Gamma distribution, by a finite step distribution. One may conclude that there are, say, k components in the model, but obviously this is irrational. If the true mixing distribution is very skew we expect that the support points will be separated, some near the positive axis but surely some points will be at the right tail. Can we draw conclusions about the number of components?

Investigation of this issue is still under research. The main difficulty in examining this problem is simply that there does not exist an algorithm which can easily give the SML estimate with a few support points.

However, if we relax the conditions for the number of support points then things are more clear. The algorithms will converge, and a mixing distribution can be obtained. But this idea suffers from the fact that the likelihood cannot increase when redundant points are added, and, hence, the derived estimate of the mixing distribution is inconsistent.

However, when the number of support points is large, a more smoothed estimate of the mixing distribution is obtained for further use. For example, we can use it for empirical Bayes purposes, (see Laird (1982)). Another idea is to use further methods to smooth it like a kernel density estimation method.

Moreover, this smooth estimate of the mixing distribution can be used for testing hypotheses. The idea stems from the fact that for discrete data goodness of fit tests are usually asymptotic and thus they lack power. For continuous cases the tests for goodness of fit are more powerful and thus they are more trustworthy.

Many problems remain open and a lot of research must be carried out before these problems can ultimately be answered. Note however, that Chen (1995) and McKay (1996) showed that any estimator of the mixing distribution converges very slowly to the true mixing distribution and the loss of information is great. These results reveal that we need a large sample size in order to approximate satisfactorily the true mixing distribution.

Chapter 4

Other Estimation Methods For Finite Mixtures

4.1 Introduction

In the previous chapter, we described the ML method of estimation for finite mixtures. The easily programmable EM algorithm has led to the wide acceptance of the ML method as a method of estimation for finite mixture models in a vast number of applications. Several other methods have been proposed for estimation purposes, but they failed to be widely used, mainly because of problems in applying them and secondly because of the fact that their properties are not well understood.

The method of moments (hereafter MM) is a well known counterpart to the ML method. Historically, the first attempt for estimation in normal mixtures was made using the moment method (Pearson, 1894). In this chapter, the MM for finite Poisson mixtures is presented, along with a general review of the method for general mixtures. The efficiency of the method relative to the ML method is examined. Another important (but rather overlooked) issue that is examined concerns the existence of the moment estimates. By the term 'existence' we refer to the cases where the moment equations lead to estimates that are in the admissible range.

In the sequel, a variant of the method of moments is proposed. This method utilises the zero sample frequency instead of the third sample moment and so, for data sets with a high zero frequency, it can be a more appropriate method of estimation. Recall, that the third moment has a great variability which can significantly affect the stability of the derived estimators.

The last section of this chapter is devoted to the Bayesian approach of estimation in finite mixtures. Here, again, a lot of computational problems in deriving posterior

distributions has resulted in a low applicability of the method. Recent developments, however, that have been considered in the Bayesian approach of estimation problems, via the well known Gibbs sampler and its variants, facilitate the use of Bayesian estimation techniques.

4.2 The method of moments

When applying the MM to finite mixtures, one has to distinguish between two different approaches:

- The first applies to a k-finite mixture (with a known number of components) and equates the first m moments of the hypothesised distribution to the sample moments, as it is the case with the usual MM. The number m depends on the number k of components. For example, for a k-finite Poisson mixture one needs $m = 2k - 1$ equations.
- The second refers to the case where the number k of support points is unknown prior to data investigation and is known as a semiparametric case. This case reduces to the successive application of the MM for each fixed value of k.

A brief review of both of these approaches is provided in the next sections.

4.2.1 Moment Estimation with Known k

The ÌÌ was the first method employed for estimation in finite mixture problems. Pearson (1894) tried to estimate a 2-finite normal mixture by equating the population moments to the sample moments. He obtained a nonic (9th degree) equation. Solving this system of equations he obtained estimates for the parameters under consideration which, however, were not unique. So, he chose the solution whose sixth moment was closer to the observed one.

Clearly, such an approach is laborious. In the years that followed the interest was concentrated on relaxing the complexity of such a solution and many authors tried to propose strategies for reducing the effort (see, e.g., Cohen, 1967). A comprehensive account of such attempts up to 1980 was given by Gupta and Huang (1981). In recent years the impact of high-speed computers, which can solve non-linear systems of equations rapidly, has limited the interest to special numerical methods needed for applying the MM.

Except for the case of normal mixtures described above, Rider (1961, 1962) treated several other mixtures, including the binomial, the Poisson and the exponential cases. Rider

(1961) obtained the solution for 2-finite mixture of Poisson distributions and he showed that the moment estimators are consistent. Blischke (1964, 1965) treated the case of binomial mixtures. John (1970) derived moment estimators and their asymptotic distributions for 2-finite mixtures of binomial, Poisson, negative binomial and hypergeometric distributions.

Tan and Chang (1972) compared the MM to the ML for a 2-finite mixture of normal distributions. Their findings suggested that the ML is superior almost always, especially when the components are well determined. For components close together, the gain in efficiency of the ML method is very low, while the computational effort required is much increased.

The MM is known to have some disadvantages. For example, moments of high order have large variances and hence they are not very suitable for estimating purposes. In order to overcome these disadvantages some modifications of the method can be proposed. These pertain to replacing moments of high order with other functionals that have lower variances.

Kabir (1968) proposed another estimation method for the case of finite mixtures from the exponential family which is a generalisation of the method of moments. He used functionals of the observed probabilities. This procedure estimates separately the mixing proportions from the remaining parameters, so that the method has two distinct steps. Because of its complexity, this method has not attracted a lot of attention. Kabir showed the asymptotic normality of the derived estimators. The performance of the method, however, has not been investigated. Redner and Walker (1984) described this method as a generalised method of moments.

Tallis and Light (1968) proposed the use of fractional moments. They showed that, for exponential mixtures, the efficiency of the estimators can be increased considerably by choosing appropriate fractional moments. Unfortunately, the applicability of the use of fractional moments is limited by their complexity. For example, finding the fractional moments of the exponential distribution is quite simple but is a tedious task in the case of the Poisson distribution.

The MM has also been utilised in the case of multivariate mixtures, mainly multivariate normal mixtures. Day (1969) discussed the MM for a 2-finite multivariate normal mixture. The large number of equations that need to be solved simultaneously, makes the method almost inapplicable. However, Lindsay and Basak (1993) considered another system of equations which facilitates the estimation by choosing moments that reduce the complexity of the derived system of equations.

In the present chapter, the efficiency of the MM relative to the ML method is examined for 2-finite Poisson mixtures. This study contains both a comparison of the asymptotic efficiency of the two methods, based on their asymptotic variance-covariance matrices, and a small sample comparison via simulation. The latter approach is of greater practical interest, since the asymptotic results hold only for quite large sample sizes. This simulation study covers another important, but rather undertreated, aspect of the MM. For small sample sizes, the moment estimates may not exist, since the system of estimating equation may not have a solution. This aspect cannot be examined through a study of the asymptotic efficiency of the method. For example, it is known that the ML method leads to estimates with large variances if the components are close together (Hasselblad, 1969). This problem might have led to the use of moment estimates for such cases in order to avoid these large variances. Our simulation study reveals that in such cases the ML estimates suffer from large variances, but the moment estimates do not even exist.

In the later part of this chapter, a new method is proposed. This method considers replacing the third moment by the zero frequency leading to equating the observed frequency of the value 0 to its expected frequency under a 2-finite Poisson mixture. The method is shown to have an increased efficiency when the zero frequency is high.

4.2.2 The Method of Moments when k is Unknown : The Semiparametric Case

Recall that the term semiparametric refers to the case where the number of support points is not known a priori and the value of k must be estimated from the data.

Tucker (1963) initiated the use of this approach by considering a certain procedure to estimate the mixing distribution G of a mixed Poisson distribution via the method of moments. Recall that for the case of mixed Poisson distributions we are able to consistently estimate the moments of the mixing distribution from the sample moments (see section 2.2.2). This reduces the problem of estimating the mixing distribution to a problem of determining a distribution with given moments. This is the well known moment problem (see, e.g., Shohat and Tamarkin, 1943). Tucker (1963) proposed to estimate the moments of the mixing distribution from the data and then to solve a moment problem, so as to obtain the estimated distribution for the mixing distribution. Tucker (1963) proposed starting with k=1 and keep adding support points until the moment problem becomes intractable.

Problems connected with the existence of such a solution inhibit the use of the method. Rolph (1968), following Tucker's method proposed a method of moment estimation for a mixing distribution in (0,1) which is applicable to the binomial case. Brockett (1977) generalised Tucker's method and showed the consistency of the method considered.

A further generalisation of the methods was given by Lindsay (1989). He showed that, for several distributions belonging to the quadratic variance natural exponential family of Morris (1982), a consistent estimator of some functionals of the distribution can be obtained in the case of mixtures from this family. So, for these distributions, the moments of the mixing distribution can be found. He also proposed a method for determining the number of support points .

Heckman and Walker (1990) and Heckman *et al.* (1990) examined the case of exponential mixtures. Heckman (1990) treated the geometric mixture case, too. The main problem of such approaches remains the restricted number of components which can be estimated. For example, Heckman *et al.*. (1990), working with the exponential distribution, reported that it is very common that the moment problem has no solution for more than two moments and, hence, this restricts the applicability of the approach. Withers (1991, 1996) examined the case of moments estimators for some families of mixture distributions.

The method of moments is also popular for estimating the mixing distribution in the context of the empirical Bayesian approach. This is taken up later in section 4.7.

A different approach was proposed by Rutherford and Krutchkoff (1967). They proposed the use of the first four sample moments for choosing the member of the Pearson' s family which possesses these first four moments. The advantage of such an approach is that we estimate the mixing distribution from a broad family of continuous distributions (the Pearson's family) and hence we always obtain a smooth mixing distribution. Unfortunately, the practical use of the method is doubtful for two main reasons. The first is that we need four moments and the moment problem may be intractable. The second is that the estimated member of the Pearson family can be defined over the entire real line which clearly contradicts the case of Poisson mixtures whose parameters ought to be positive. However, they showed that such an estimator is consistent and converges almost surely to the true mixing distribution.

4.2.3 Critique of the Method of Moments

The MM has certain disadvantages. Theoretically, if the data truly come from a mixture, the moment problem is expected to be solvable, especially when the sample size is not small. Failure to solve the moment problem with real data, might be considered as an indication of departure from the assumed model. However, this is an erroneous conclusion, because, as shown later in 4.3, even in cases where the model is correctly specified and the sample size is moderately large, the moment problem may be intractable. For example, consider the case where we sample from a 2-finite Poisson mixture. It is known that for such a model the variance must be larger than the mean, but there exists a high probability for this not to happen due to random variability, especially when the two components are close together.

On the other hand, it would not be appropriate to use a large number of estimated moments because of the high variability of high order sample moments. We need 2k-1 moments in order to estimate a distribution with k support points. So, using the first three moments one can estimate only a 2-finite step distribution (since there will be three equations with three unknowns) and the estimated distribution will be very sparse. Using more moments results in estimators that are not efficient due to the high variability of the higher order sample moments .

In order to overcome these difficulties one can arbitrarily assume that the mixing distribution is a k-finite step distribution with equal probabilities at k distinct points (see, e.g., Maritz and Lwin, 1989), and then solve the relevant equations. In doing so, one can use the first k moments to obtain k distinct points with non zero probability masses or do the opposite i.e. choose k equally distanced points and solve for the corresponding probabilities.

It must be emphasised that the efficiency and the plausibility of such estimation procedures, with some relaxation in favour of simplicity, depend on what one really wishes to do. For example, for the purpose of empirical Bayesian problems a rough estimation of the distribution may be sufficient as Maritz (1989) pointed out.

Maritz (1969), in a paper for empirical Bayes estimation for the Poisson distribution, proposed a method where the two first moments are equated to their observed counterparts while the absolute distance between the observed and the theoretical third moments is minimised.

Note that the method proposed by Quandt and Ramsey (1978) using the moment generating function is a variant of the MM. They did not equate theoretical to observed

moments but they required them to be as close as possible, using weights that were based on the order of moments, so as higher order moments were given less weight. So, higher order moments are less reliable for the estimation.

The next two sections treat the questions raised earlier concerning the non-existence and small efficiency aspects. As far as the non-existence is concerned we have not considered any remedies when the moment equations fail to lead to valid estimates. Our findings focus on the classical moment method that equates the theoretical moments to their sample counterparts. It remains an open problem to examine thoroughly variants of the moment method that overcome the 'existence' problems examined in this thesis.

4.3 Existence of the Moment Estimates

As mentioned earlier, the existence of the moment estimates is not certain for small or moderate sample sizes even when the model is correctly specified. This fact reduces the practicability of the MM because, in practice, sample sizes rarely are sufficiently large.

Shohat and Tamarkin (1943) treated the moment problem in depth. They provided necessary conditions for finding a distribution with a given finite series of moments $(\mu_1, \mu_2, \ldots, \mu_s)$. For a distribution with support in (0,+∞), the conditions are

$$\begin{vmatrix} 1 & \mu_1 & \ldots & \mu_s \\ \mu_1 & \mu_2 & \ldots & \mu_{s+1} \\ . & . & \ldots & . \\ \mu_s & \mu_{s+1} & \ldots & \mu_{2s} \end{vmatrix} \geq 0 \qquad \text{and} \qquad \begin{vmatrix} \mu_1 & \mu_2 & \ldots & \mu_{s+1} \\ \mu_2 & \mu_3 & \ldots & \mu_{s+2} \\ . & . & \ldots & . \\ \mu_{s+1} & \mu_{s+2} & \ldots & \mu_{2s+1} \end{vmatrix} \geq 0$$

For the first five moments the conditions are:

$$\begin{aligned} &\mu_1 \geq 0 \\ &\mu_2 \geq \mu_1^2 \\ &\mu_1\mu_3 \geq \mu_2^2 \\ &\mu_2\mu_4 - \mu_1^2\mu_4 - \mu_2^3 - \mu_3^2 + 2\mu_1\mu_2\mu_3 \geq 0 \\ &\mu_1\mu_3\mu_5 - \mu_1\mu_4^2 - \mu_2^2\mu_5 - \mu_3^2 + 2\mu_2\mu_3\mu_4 \geq 0 \end{aligned} \qquad . \qquad (4.1)$$

In order to examine how many support points one is able to estimate, a simulation experiment was carried out. From some mixed Poisson distribution, samples of varying size were generated, for a variety of parameters values. Applying the conditions given in (4.1) the number of support points which are estimable by the MM was subsequently derived. The

distributions used were 2-finite Poisson mixtures, 3-finite Poisson mixtures and the negative binomial distribution as given in (2.22) with parameters $a=(1-p)/p, b$ for various choices of b and p.

The relationships between the moments of the mixed Poisson distribution to those of the mixing distribution are given in section 2.2.2. Solving the system of estimating equations we find that the moments of the mixing distribution are related to the moments of the mixed Poisson distribution by the following formulae

$$\begin{aligned} E(\lambda) &= E(X) \\ E(\lambda^2) &= E(X^2) - E(X) \\ E(\lambda^3) &= E(X^3) - 3E(X^2) + 2E(X) \\ E(\lambda^4) &= E(X^4) - 6E(X^3) + 11E(X^2) - 6E(X) \\ E(\lambda^5) &= E(X^5) - 10E(X^4) + 35E(X^3) - 50E(X^2) + 24E(X) \end{aligned} \tag{4.2}$$

In practical situations the moments of X are replaced by the corresponding sample moments as defined by

$$\mu_r = \frac{\sum_{i=1}^{n} X_i^r}{n} \quad . \tag{4.3}$$

For the simulation experiment, the moments of the mixing distribution were estimated from the sample using (4.2) and substituting the theoretical moments $E(X^r)$ with their sample counterparts given in (4.3). Then the validity of the conditions given in (4.1) was checked. The number of support points which were estimable was calculated for each sample. Tables 4.1-4.3 summarise the results of this simulation experiment. Their entries represent the proportion of times the MM failed in estimating a solution with k support points (k=2,3) in 10000 replications.

It is evident from the results that the obtained estimates have a few support points. For example, looking at Table 4.2, one can see that for small to moderate sample sizes the MM fails to give 2 support points, even though we sampled from a 2-finite distribution. The same is true in Table 4.3 for 3- finite Poisson mixtures. Even for large sample sizes the method fails to reconstruct the true mixing distribution with 3 support points. For the negative binomial distribution, it can be seen from Table 4.1, that as the overdispersion decreases the method of moments fails. It is interesting that when the mixing distribution is a continuous distributions, (as in the case of the negative binomial distribution) it is possible to estimate only a few

support points using the MM. This is true even when the sample size is relatively large. Note also that the moment estimates of the parameters of the negative binomial distribution very often do not exist (see Johnson *et al.*, 1992). Recall that in chapter 3, when examining the choice of initial values for the EM algorithm, it was shown that the moment estimates did not exist with high probability.

Table 4.1
Proportion of times the moment method failed to give estimates with k support points in 10000 replications. Data were generated from a negative binomial distribution with parameters $a = (1-p)/p, b$.

sample size		**50**		**100**		**250**		**500**	
number of support points		**2**	**3**	**2**	**3**	**2**	**3**	**2**	**3**
b	**p**								
	0.05	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00
	0.10	0.01	0.33	0.00	0.07	0.00	0.00	0.00	0.00
	0.15	0.01	0.69	0.00	0.31	0.00	0.05	0.00	0.00
	0.20	0.00	0.92	0.01	0.59	0.00	0.19	0.00	0.04
	0.25	0.03	0.98	0.01	0.82	0.00	0.41	0.00	0.16
	0.30	0.08	1.00	0.01	0.94	0.00	0.61	0.00	0.34
	0.35	0.15	1.00	0.03	0.99	0.00	0.79	0.00	0.54
	0.40	0.23	1.00	0.08	1.00	0.00	0.91	0.00	0.72
	0.45	0.33	1.00	0.14	1.00	0.02	0.98	0.01	0.87
0.50	0.50	0.43	1.00	0.22	1.00	0.04	0.99	0.00	0.96
	0.55	0.53	1.00	0.33	1.00	0.12	1.00	0.03	0.99
	0.60	0.61	1.00	0.42	1.00	0.19	1.00	0.06	1.00
	0.65	0.70	1.00	0.52	1.00	0.29	1.00	0.14	1.00
	0.70	0.78	1.00	0.63	1.00	0.39	1.00	0.23	1.00
	0.75	0.86	1.00	0.73	1.00	0.50	1.00	0.35	1.00
	0.80	0.94	1.00	0.85	1.00	0.66	1.00	0.54	1.00
	0.85	0.95	0.98	0.92	1.00	0.79	1.00	0.66	1.00
	0.90	0.93	0.93	1.00	1.00	1.00	1.00	1.00	1.00
	0.95	0.73	0.73	0.92	0.92	1.00	1.00	1.00	1.00
	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.10	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00
	0.15	0.00	0.31	0.00	0.07	0.00	0.00	0.00	0.00
	0.20	0.00	0.57	0.00	0.25	0.00	0.04	0.00	0.00
	0.25	0.00	0.78	0.01	0.47	0.00	0.14	0.00	0.03
	0.30	0.00	0.92	0.00	0.67	0.00	0.31	0.00	0.11
	0.35	0.04	0.98	0.00	0.83	0.00	0.49	0.01	0.25
	0.40	0.08	1.00	0.01	0.93	0.00	0.68	0.00	0.44
	0.45	0.16	1.00	0.04	0.98	0.00	0.81	0.00	0.61
1.00	0.50	0.24	1.00	0.09	1.00	0.01	0.92	0.01	0.79
	0.55	0.33	1.00	0.15	1.00	0.02	0.96	0.00	0.88
	0.60	0.42	1.00	0.24	1.00	0.07	0.99	0.01	0.94
	0.65	0.52	1.00	0.34	1.00	0.14	1.00	0.04	0.98
	0.70	0.63	1.00	0.44	1.00	0.23	1.00	0.10	1.00
	0.75	0.70	1.00	0.55	1.00	0.35	1.00	0.20	1.00
	0.80	0.80	1.00	0.66	1.00	0.48	1.00	0.33	1.00
	0.85	0.90	1.00	0.80	1.00	0.59	1.00	0.46	1.00
	0.90	0.96	0.99	0.92	1.00	0.81	1.00	0.67	1.00
	0.95	0.92	0.92	0.99	0.99	1.00	1.00	1.00	1.00

Table 4.1 (continued)

sample size		50		100		250		500	
number of support points		2	3	2	3	2	3	2	3
b	p								
	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.15	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00
	0.20	0.01	0.14	0.00	0.02	0.00	0.00	0.00	0.00
	0.25	0.00	0.31	0.00	0.10	0.00	0.00	0.00	0.00
	0.30	0.00	0.49	0.00	0.23	0.00	0.04	0.00	0.00
	0.35	0.01	0.65	0.01	0.40	0.01	0.12	0.00	0.02
	0.40	0.01	0.79	0.01	0.56	0.01	0.25	0.00	0.09
	0.45	0.02	0.89	0.00	0.69	0.00	0.40	0.01	0.21
3.00	0.50	0.05	0.95	0.00	0.81	0.01	0.55	0.00	0.35
	0.55	0.10	0.98	0.02	0.90	0.01	0.68	0.00	0.50
	0.60	0.18	1.00	0.06	0.96	0.00	0.80	0.01	0.65
	0.65	0.27	1.00	0.11	0.98	0.01	0.90	0.00	0.81
	0.70	0.38	1.00	0.22	1.00	0.05	0.95	0.00	0.88
	0.75	0.49	1.00	0.32	1.00	0.13	0.99	0.03	0.96
	0.80	0.59	1.00	0.45	1.00	0.25	1.00	0.13	0.99
	0.85	0.71	1.00	0.58	1.00	0.41	1.00	0.27	1.00
	0.90	0.80	1.00	0.70	1.00	0.58	1.00	0.46	1.00
	0.95	0.96	1.00	0.91	1.00	0.79	1.00	0.67	1.00
	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.15	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	0.20	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00
	0.25	0.01	0.17	0.01	0.04	0.00	0.00	0.00	0.00
	0.30	0.00	0.34	0.01	0.11	0.00	0.01	0.00	0.00
	0.35	0.00	0.48	0.00	0.24	0.00	0.04	0.00	0.01
	0.40	0.00	0.64	0.00	0.40	0.01	0.13	0.00	0.03
	0.45	0.01	0.76	0.00	0.53	0.00	0.26	0.00	0.10
5.00	0.50	0.02	0.86	0.00	0.67	0.01	0.39	0.00	0.20
	0.55	0.05	0.93	0.00	0.78	0.01	0.54	0.00	0.35
	0.60	0.11	0.97	0.03	0.87	0.00	0.66	0.00	0.50
	0.65	0.19	0.99	0.06	0.93	0.01	0.79	0.01	0.65
	0.70	0.29	1.00	0.12	0.97	0.02	0.88	0.00	0.78
	0.75	0.39	1.00	0.23	0.99	0.08	0.95	0.01	0.91
	0.80	0.51	1.00	0.37	1.00	0.16	0.98	0.06	0.94
	0.85	0.63	1.00	0.50	1.00	0.32	1.00	0.18	0.99
	0.90	0.75	1.00	0.66	1.00	0.50	1.00	0.38	1.00
	0.95	0.89	1.00	0.81	1.00	0.74	1.00	0.72	1.00

Table 4.2
Proportion of times the moment method failed to give estimates with k support points in 10000 replications. Data were generated from a 2-finite Poisson mixture with parameters p_1, $ë_1$ and $ë_2$

	sample size		50		100		250		500	
			number of support points							
p_1	$ë_1$	$ë_2$	2	3	2	3	2	3	2	3
0.10	1	2	0.71	1.00	0.64	1.00	0.56	0.99	0.52	0.98
	1	3	0.62	1.00	0.55	0.99	0.43	0.96	0.35	0.92
	1	4	0.55	0.99	0.48	0.97	0.36	0.91	0.28	0.85
	1	5	0.51	0.97	0.44	0.94	0.32	0.85	0.23	0.78
	2	3	0.69	1.00	0.63	0.99	0.57	0.98	0.53	0.98
	2	4	0.59	0.99	0.51	0.98	0.41	0.95	0.32	0.91
	2	5	0.52	0.98	0.44	0.96	0.31	0.89	0.20	0.83
	3	4	0.68	1.00	0.64	0.99	0.59	0.98	0.54	0.97
	3	5	0.59	0.99	0.53	0.98	0.43	0.95	0.35	0.92
	4	5	0.66	0.99	0.64	0.99	0.58	0.97	0.54	0.97
0.30	1	2	0.63	1.00	0.54	1.00	0.39	0.98	0.30	0.96
	1	3	0.45	1.00	0.31	0.97	0.15	0.90	0.07	0.83
	1	4	0.32	0.97	0.20	0.91	0.07	0.82	0.01	0.76
	1	5	0.26	0.94	0.15	0.86	0.03	0.77	0.01	0.73
	2	3	0.63	1.00	0.55	0.99	0.44	0.98	0.32	0.95
	2	4	0.43	0.99	0.29	0.96	0.14	0.90	0.04	0.83
	2	5	0.27	0.97	0.14	0.90	0.03	0.82	0.00	0.76
	3	4	0.63	0.99	0.56	0.99	0.46	0.97	0.38	0.95
	3	5	0.45	0.99	0.32	0.96	0.16	0.90	0.06	0.85
	4	5	0.62	0.99	0.57	0.98	0.48	0.96	0.40	0.94
0.50	1	2	0.58	1.00	0.47	1.00	0.28	0.98	0.16	0.95
	1	3	0.31	1.00	0.16	0.97	0.04	0.89	0.00	0.82
	1	4	0.16	0.98	0.04	0.91	0.00	0.82	0.01	0.78
	1	5	0.08	0.94	0.01	0.85	0.00	0.77	0.00	0.74
	2	3	0.60	1.00	0.51	0.99	0.36	0.98	0.22	0.94
	2	4	0.31	0.99	0.16	0.96	0.05	0.90	0.00	0.85
	2	5	0.12	0.97	0.03	0.91	0.00	0.82	0.00	0.76
	3	4	0.61	1.00	0.54	0.99	0.41	0.97	0.30	0.95
	3	5	0.36	0.99	0.21	0.96	0.05	0.89	0.01	0.85
	4	5	0.60	0.99	0.54	0.98	0.44	0.96	0.34	0.94
0.70	1	2	0.58	1.00	0.45	1.00	0.27	0.99	0.13	0.96
	1	3	0.23	1.00	0.08	0.98	0.00	0.92	0.00	0.87
	1	4	0.05	0.99	0.00	0.94	0.01	0.85	0.00	0.79
	1	5	0.01	0.97	0.00	0.89	0.00	0.81	0.00	0.76
	2	3	0.61	1.00	0.51	0.99	0.37	0.98	0.24	0.96
	2	4	0.30	1.00	0.13	0.97	0.02	0.92	0.01	0.87
	2	5	0.08	0.98	0.01	0.93	0.00	0.85	0.00	0.80
	3	4	0.61	1.00	0.54	0.99	0.42	0.97	0.32	0.95
	3	5	0.35	0.99	0.21	0.97	0.05	0.91	0.01	0.87
	4	5	0.61	0.99	0.56	0.98	0.45	0.97	0.38	0.95
0.90	1	2	0.67	1.00	0.59	1.00	0.44	1.00	0.29	0.99
	1	3	0.39	1.00	0.22	1.00	0.05	0.99	0.00	0.96
	1	4	0.18	1.00	0.05	1.00	0.00	0.95	0.01	0.90
	1	5	0.06	1.00	0.01	0.98	0.00	0.91	0.01	0.84
	2	3	0.68	1.00	0.61	1.00	0.51	0.99	0.43	0.98
	2	4	0.49	1.00	0.34	0.99	0.14	0.97	0.04	0.94
	2	5	0.27	1.00	0.11	0.99	0.01	0.94	0.01	0.90
	3	4	0.68	1.00	0.62	0.99	0.56	0.98	0.49	0.97
	3	5	0.52	1.00	0.42	0.99	0.23	0.96	0.11	0.93
	4	5	0.68	1.00	0.63	0.99	0.57	0.98	0.52	0.97

Table 4.3
Proportion of times the moment method failed to give estimates with k support points in 10000 replications. Data were generated from a 3-finite Poisson mixture with parameters p_1,p_2, $ë_1$,$ë_2$ and $ë_3$

					sample size		50		100		250		500	
					number of support points									
p_1	p_2	$ë_1$	$ë_2$	$ë_3$										
					2	3	2	3	2	3	2	3		
0.25	0.2	1	2	6	0.07	0.90	0.02	0.82	0.00	0.73	0.01	0.69		
			4	6	0.17	0.90	0.08	0.82	0.00	0.70	0.00	0.63		
		2	2	6	0.05	0.91	0.01	0.85	0.00	0.77	0.00	0.72		
			4	6	0.16	0.94	0.06	0.86	0.00	0.75	0.01	0.69		
		3	2	6	0.10	0.93	0.02	0.87	0.00	0.77	0.00	0.72		
			4	6	0.25	0.97	0.11	0.91	0.01	0.82	0.01	0.76		
	0.4	1	2	4	0.18	0.99	0.07	0.95	0.01	0.85	0.00	0.78		
			6	4	0.15	0.92	0.05	0.82	0.01	0.69	0.00	0.59		
		2	2	4	0.28	0.99	0.14	0.97	0.03	0.91	0.00	0.86		
			6	4	0.14	0.95	0.04	0.87	0.00	0.76	0.00	0.67		
		3	2	4	0.39	0.99	0.23	0.97	0.08	0.92	0.01	0.85		
			6	4	0.23	0.97	0.11	0.93	0.02	0.85	0.01	0.79		
0.50	0.2	1	4	2	0.14	1.00	0.04	0.97	0.00	0.87	0.01	0.79		
			6	2	0.00	0.97	0.00	0.89	0.00	0.78	0.00	0.71		
			4	2	0.35	1.00	0.20	0.99	0.04	0.94	0.00	0.89		
			6	2	0.02	0.98	0.00	0.93	0.00	0.85	0.01	0.80		
			4	2	0.50	1.00	0.35	0.98	0.18	0.94	0.08	0.89		
			6	2	0.08	0.98	0.02	0.93	0.01	0.83	0.00	0.75		
	0.4	1	2	6	0.04	1.00	0.00	0.95	0.01	0.84	0.00	0.75		
			4	6	0.06	0.94	0.00	0.83	0.00	0.68	0.00	0.56		
		2	2	6	0.13	1.00	0.03	0.98	0.00	0.92	0.00	0.87		
			4	6	0.13	0.98	0.03	0.92	0.00	0.79	0.00	0.69		
		3	2	6	0.19	0.99	0.07	0.96	0.00	0.87	0.00	0.79		
			4	6	0.38	0.99	0.24	0.97	0.06	0.91	0.01	0.85		

Similar results were reported in Heckman and Walker (1990) for the exponential distribution. They showed that using the semiparametric method for estimating the mixing distribution in the case of mixtures of the exponential distribution, the number of estimable support points is very small.

From the above experiment it is evident that the MM fails to provide estimates very often. This can cause a lot of problems in practice. In the next section the efficiency of the method will be examined in connection with the results of this section concerning the non-existence of the moment estimates.

4.4 The Efficiency of the Moment Method for 2-Finite Poisson Mixtures

The aim of this section is to study the efficiency of the MM for 2-finite Poisson mixtures. The results will be based on both asymptotic arguments and small sample size

comparisons using simulation. The latter approach can reveal interesting evidence about the method since, as shown in the previous section, for small samples, it is possible to fail to obtain moment estimates. Thus, a comparison based merely on asymptotic arguments, when such problems are ignored, cannot reflect the properties of the estimators in practical situations.

For the case of a 2-finite Poisson mixture, the three parameters can be estimated using the first three moments of the data. The system of equations is the following

$$\left.\begin{array}{c} p_1\lambda_1 + p_2\lambda_2 = \mu_1 \\ p_1({\lambda_1}^2 + \lambda_1) + p_2({\lambda_2}^2 + \lambda_2) = \mu_2 \\ p_1({\lambda_1}^3 + 3{\lambda_1}^2 + \lambda_1) + p_2({\lambda_2}^3 + 3\lambda^2{}_2 + \lambda_2) = \mu_3 \end{array}\right\} \tag{4.4}$$

where $ì_k$, *k=1, 2, 3,* are the sample moments. Solving this system of equations we obtain

$$\hat{\lambda}_1, \hat{\lambda}_2 = \frac{-b \pm \sqrt{D}}{2a},$$

where $b = (\mu_3 - 3\mu_2 + 2\mu_1 - \mu_1\mu_2 + \mu_1^2)$, $a = ({\mu_1}^2 - \mu_2 + \mu_1)$ and $D = b^2 - 4a\left({\mu_2}^2 - {\mu_1}^2 + \mu_1\mu_2 - \mu_1\mu_3\right)$.

Since it must hold that $\hat{\lambda}_1 < \hat{\lambda}_2$ the estimate for ë$_1$ is the smaller root. Then we obtain

$$\hat{p}_1 = \frac{\mu_1 - \hat{\lambda}_1}{\hat{\lambda}_2 - \hat{\lambda}_1} \quad .$$

The asymptotic variance-covariance matrix of the estimates can be calculated as follows. Using first order Taylor expansion (see, e.g., Titterington *et al.*, 1985) we obtain that the asymptotic variance covariance matrix **V** is calculated as $\mathbf{V= G^{-1}MG^{-1}}$ where

$$\mathbf{G} = \begin{bmatrix} \lambda_1 - \lambda_2 & p_1 & p_2 \\ {\lambda_1}^2 + \lambda_1 - \left({\lambda_2}^2 + \lambda_2\right) & p_1\left(2\lambda_1 + 1\right) & p_2\left(2\lambda_2 + 1\right) \\ {\lambda_1}^3 + 3{\lambda_1}^2 + \lambda_1 - \left({\lambda_2}^3 + 3{\lambda_2}^2 + \lambda_2\right) & p_1\left(3{\lambda_1}^2 + 6\lambda_1 + 1\right) & p_2\left(3{\lambda_2}^2 + 6\lambda_2 + 1\right) \end{bmatrix}$$

and **M** is the 3×3 matrix with its ij-th element representing the covariance between the i-th and j-th sample moments. Following Stuart and Ord (1994) the elements M_{ij} of **M** are

$$M_{ij} = \frac{1}{n}\left(E(X^{i+j}) - E(X^i)E(X^j)\right).$$

For a 2-finite Poisson mixture the first six simple moments are given by the following formulas

$$\begin{aligned}
E(X) &= p_1\lambda_1 + p_2\lambda_2 \\
E(X^2) &= p_1\left(\lambda_1^2 + \lambda_1\right) + p_2\left(\lambda_2^2 + \lambda_2\right) \\
E(X^3) &= p_1\left(\lambda_1^3 + 3\lambda_1^2 + \lambda_1\right) + p_2\left(\lambda_2^3 + 3\lambda_2^2 + \lambda_2\right) \\
E(X^4) &= p_1\left(\lambda_1^4 + 6\lambda_1^3 + 7\lambda_1^2 + \lambda_1\right) + p_2\left(\lambda_2^4 + 6\lambda_2^3 + 7\lambda_2^2 + \lambda_2\right) \\
E(X^5) &= p_1\left(\lambda_1^5 + 10\lambda_1^4 + 25\lambda_1^3 + 15\lambda_1^2 + \lambda_1\right) + p_2\left(\lambda_2^5 + 10\lambda_2^4 + 25\lambda_2^3 + 15\lambda_2^2 + \lambda_2\right) \\
E(X^6) &= p_1\left(\lambda_1^6 + 15\lambda_1^5 + 65\lambda_1^4 + 90\lambda_1^3 + 31\lambda_1^2 + \lambda_1\right) + \\
&\quad + p_2\left(\lambda_2^6 + 15\lambda_2^5 + 65\lambda_2^4 + 90\lambda_2^3 + 31\lambda_2^2 + \lambda_2\right).
\end{aligned}$$

Respectively the variance-covariance matrix of the ML estimates was given in section 3.2.3.

Example 3.1 (continued) Consider the data of Example 3.1 given in Table 3.1. Assuming a 2-finite Poisson model the obtained moment estimates of the parameters were p_1=0.697 (0.204), ë$_1$=1.537 (0.324) and ë$_2$=3.863 (0.816). We can see that they differ only slightly from the ML estimates and that the jacknife standard errors calculated are much larger for the moment estimates.

The asymptotic efficiencies of the MM are reported in Tables 4.4a-4.4c for several combinations of parameter values. The entries of tables are the values of the ratio $|V_{ML}|/|V_{MM}|$, where $|V|$ denotes the generalised variance, and the subscripts indicate the method used. Entries lower than 1 favour the ML method. Ét can be seen that, for well separated components, the efficiency is low due to the low variances of the ML methods. For components close together, the efficiency is higher, especially for mixing proportions near 0.5.

Figure 4.1 depicts the asymptotic efficiency of the moment method for ë$_1$=1 and various choices of the mixing proportion and the second parameter. Clearly, the efficiency

decreases rapidly as the second component gets further and further away from the first component.

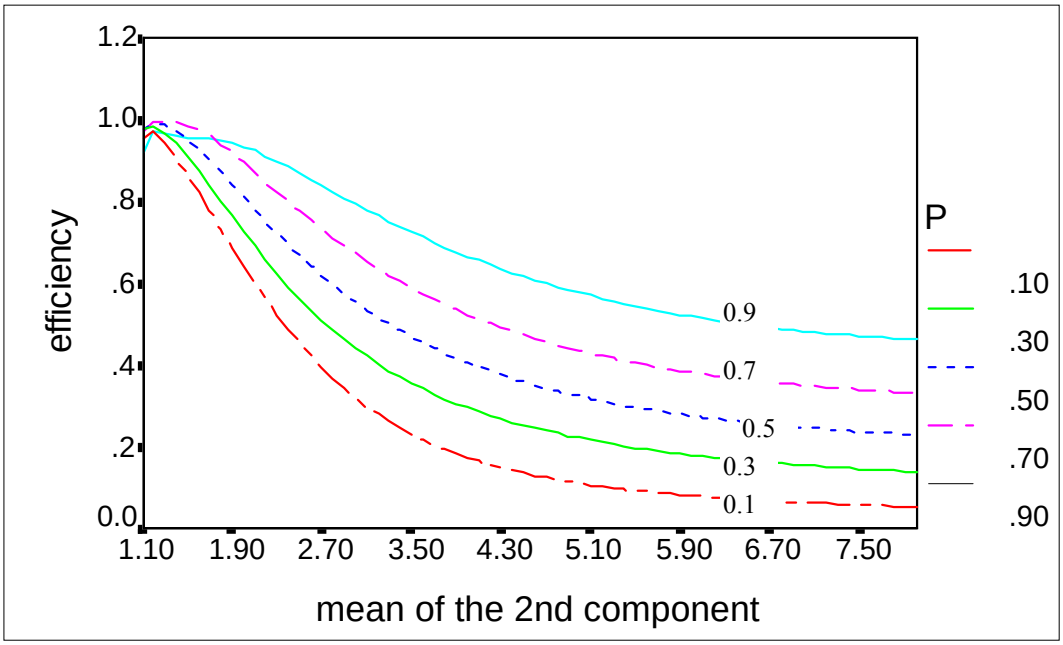


Figure 4.1 Asymptotic efficiency of the method of moments for 2-finite Poisson mixtures with $\ddot{e}_1=1$

Table 4.4a
Asymptotic efficiency of the Moment method, relative to the Maximum Likelihood method, for $\ddot{e}_1=1$

	$\ddot{e}_2$								
p_1	2	3	4	5	6	7	8	9	10
0.1	0.645	0.322	0.175	0.112	0.082	0.065	0.055	0.049	0.045
0.2	0.687	0.384	0.239	0.170	0.133	0.112	0.099	0.090	0.085
0.3	0.729	0.443	0.297	0.223	0.181	0.157	0.142	0.132	0.126
0.4	0.771	0.499	0.353	0.274	0.229	0.202	0.185	0.175	0.169
0.5	0.814	0.556	0.408	0.325	0.278	0.249	0.231	0.221	0.216
0.6	0.856	0.614	0.465	0.379	0.329	0.299	0.281	0.271	0.266
0.7	0.898	0.674	0.526	0.437	0.384	0.352	0.334	0.324	0.320
0.8	0.933	0.738	0.593	0.502	0.446	0.412	0.394	0.384	0.381
0.9	0.937	0.796	0.668	0.579	0.521	0.485	0.464	0.454	0.451

Table 4.4b
Asymptotic efficiency of the Moment method, relative to the Maximum Likelihood method, for $\ddot{e}_1=2$

	$\ddot{e}_2$								
p_1	3	4	5	6	7	8	9	10	11
0.1	0.766	0.466	0.291	0.202	0.152	0.123	0.104	0.091	0.083
0.2	0.813	0.550	0.385	0.289	0.232	0.195	0.172	0.156	0.145
0.3	0.855	0.620	0.460	0.360	0.298	0.258	0.231	0.214	0.202
0.4	0.892	0.682	0.525	0.423	0.358	0.315	0.287	0.268	0.256
0.5	0.925	0.737	0.585	0.482	0.414	0.369	0.339	0.320	0.308
0.6	0.952	0.788	0.641	0.537	0.467	0.421	0.390	0.371	0.359
0.7	0.970	0.832	0.694	0.591	0.520	0.472	0.441	0.420	0.409
0.8	0.971	0.864	0.740	0.643	0.572	0.524	0.491	0.471	0.459
0.9	0.924	0.855	0.760	0.680	0.618	0.573	0.542	0.521	0.509

Table 4.4c
Asymptotic efficiency of the Moment method, relative to the Maximum Likelihood method, for $\ddot{e}_1=3$

	$\ddot{e}_2$								
p_1	4	5	6	7	8	9	10	11	12
0.1	0.827	0.561	0.379	0.274	0.212	0.172	0.146	0.128	0.116
0.2	0.870	0.648	0.484	0.377	0.308	0.261	0.229	0.207	0.192
0.3	0.906	0.718	0.564	0.456	0.383	0.332	0.298	0.273	0.257
0.4	0.937	0.776	0.630	0.522	0.446	0.394	0.357	0.332	0.315
0.5	0.961	0.825	0.687	0.580	0.503	0.449	0.411	0.385	0.367
0.6	0.978	0.866	0.737	0.632	0.555	0.499	0.460	0.433	0.415
0.7	0.985	0.895	0.778	0.678	0.602	0.546	0.506	0.479	0.460
0.8	0.972	0.906	0.805	0.715	0.643	0.588	0.549	0.522	0.503
0.9	0.919	0.864	0.789	0.717	0.659	0.614	0.580	0.556	0.539

Asymptotic results are based on the assumption that the sample size is large. In order to examine the small sample behaviour of the two methods a simulation experiment was conducted. For several 2-finite Poisson mixtures and sample sizes, 10000 samples were drawn. For each sample, the ML estimates and the moment estimates (whenever obtainable) were derived. Samples for which the moment estimates did not exist were ignored. The reason is that we wanted to compare the two methods in practical situations, i.e. when the researcher has a dataset in hand and tries to estimate the parameters. If the moment estimates are not obtainable, the moment method is not appropriate. Note that as shown later, in chapter 6, for all the cases where the ML estimate reduces to a degenerate mixing distribution, the moment method fails to lead to admissible estimates. The generalised variances for both methods were calculated. The results are reported in Table 4.5. The entries are the values of the ratio $|V_{ML}|/|V_{MM}|$, where $|V|$ denotes the generalised variance, and the subscripts indicate the method used. Entries smaller than 1 favour the ML method.

Table 4.5
Efficiency of the moment method based on 1000 simulations

	sample size				
p_1	25	50	100	250	500
		$ë_1$=1		$ë_2$=2	
0.2	0.0011	0.0062	0.0176	0.0001	0.0072
0.5	0.0007	0.0110	0.0189	0.0143	0.1131
0.8	0.0092	0.0112	0.0044	0.0003	0.0063
		$ë_1$=1		$ë_2$=3	
0.2	0.0679	0.0801	0.0143	0.5317	0.4301
0.5	0.0017	0.1144	0.3608	0.4726	0.4413
0.8	0.0295	0.0249	0.1274	0.5111	0.4735
		$ë_1$=1		$ë_2$=5	
0.2	0.4976	0.5549	0.4923	0.2435	0.1599
0.5	0.8325	0.5242	0.3882	0.3709	0.3553
0.8	0.0957	0.6545	0.5430	0.5520	0.5348
		$ë_1$=1		$ë_2$=8	
0.2	0.4685	0.2208	0.1612	0.1299	0.1067
0.5	0.4125	0.2829	0.2871	0.2402	0.2330
0.8	0.5718	0.5572	0.4399	0.4456	0.3974
		$ë_1$=1		$ë_2$=10	
0.2	0.3300	0.1828	0.1225	0.1088	0.0887
0.5	0.3464	0.2689	0.2234	0.2206	0.2104
0.8	0.5180	0.4635	0.4043	0.3824	0.3983

The results are much different from those concerning asymptotic efficiencies. The reason is that, for small sample sizes, there is a high probability that the moment estimates do not exist, as shown in the previous section. Samples with very small overdispersion usually fail to give moment estimates. However, applying the ML method to these samples, yields estimates with high variances (see, e.g., Hasselblad, 1969). So, samples for which moment estimates of the parameters can be derived lead also to ML estimates of the parameters with low variances, making the ML method preferable.

Moreover, for many samples, especially samples of small sizes, the moment estimates were in the admissible range although they were on the boundary of admissible values, i.e. p_1 near 0 or 1 and usually a value for $\ddot{e}_2$ quite large. This resulted in high variances of the moment estimates. For example, for the case $\ddot{e}_1=1$, $\ddot{e}_2=2$, $p_1=0.5$, the estimated variance of $\ddot{e}_2$ was near 85 for the moment method and 1.9 for the ML method, because of few samples with estimated value for $\ddot{e}_2$ near 100. This indicates an unstable behaviour of the moment estimates. Finally, the ML estimates showed less bias in all cases.

4.5 The Zero Frequency Method

4.5.1 The method

Because of the high sampling variance of the third moment, one may look for some other function to replace it. A usual choice for discrete distributions is the zero relative frequency of the data set which is equated to the probability of zero under the assumed distribution (see, e.g., Kemp and Kemp, 1988). So the resulting system of equations is

$$\left.\begin{array}{c} p_1\lambda_1 + p_2\lambda_2 = \mu_1 \\ p_1({\lambda_1}^2 + \lambda_1) + p_2({\lambda_2}^2 + \lambda_2) = \mu_2 \\ p_1 \exp(-\lambda_1) + p_2 \exp(-\lambda_2) = P_0 \end{array}\right\} \quad , \qquad (4.5)$$

where P_0 is the observed proportion of zero values in the sample. For distributions with high probability at 0 this method is expected to work satisfactorily. Moreover, due to the lower variance of the proportion of zeroes relative to the third sample moment the method is expected to have a higher efficiency. System (4.5) differs from system (4.4) only in the third equation.

One can solve this system by replacing p_1 and $\ddot{e}_2$ in the third equation expressed only in terms of $\ddot{e}_1$ and then solving the 3rd equation with respect to $\ddot{e}_1$.

The resulting equation is given by

$$a\exp(-\lambda_1)+(1-a)\exp(-b)=P_0 \tag{4.6}$$

where $b=\dfrac{\mu_2-\mu_1-\mu_1\lambda_1}{\mu_1-\lambda_1}$ and $a=\dfrac{\mu_1-\lambda_1}{b-\lambda_1}$.

Equation (4.6) is non-linear and a numerical technique is required for solving it. A simple iterative scheme such as the Newton-Raphson method can be utilised. A 'good' initial choice for the value $\ddot{e}_1$ would be to use $\lambda_1^{(0)}=\ln P_0$, i.e. the value if a simple Poisson model had been assumed and then update the estimate using:

$$\lambda_1^{(i+1)}=\lambda_1^{(i)}-\frac{f(\lambda_1^{(i)})}{f'(\lambda_1^{(i)})} \tag{4.7}$$

where $f(\lambda_1)=a\exp(-\lambda_1)+(1-a)\exp(-b)-P_0$ and

$$f'(\lambda_1)=-a\exp(-\lambda_1)+\frac{\mu-b}{(b-\lambda_1)^2}\left[\exp(-\lambda_1)-\exp(-b)+1\right] .$$

Equation (4.7) is the familiar Newton-Raphson iteration.

Having obtained $\hat{\lambda}_1$, the remaining parameters are estimated by

$$\hat{\lambda}_2=\frac{\mu_2-\mu_1-\mu_1\hat{\lambda}_1}{\mu_1-\hat{\lambda}_1} \quad \text{and} \quad \hat{p}_1=\frac{\mu_1-\hat{\lambda}_1}{\hat{\lambda}_2-\hat{\lambda}_1} .$$

The asymptotic variance covariance matrix using a first order Taylor approximation is calculated as $\mathbf{V}=\mathbf{G}^{-1}\mathbf{M}\mathbf{G}^{-1}$ where

$$\mathbf{G} = \begin{bmatrix} \lambda_1 - \lambda_2 & p_1 & p_2 \\ {\lambda_1}^2 + \lambda_1 - \left({\lambda_2}^2 + \lambda_2\right) & p_1(2\lambda_1 + 1) & p_2(2\lambda_2 + 1) \\ \exp(-\lambda_1) - \exp(-\lambda_2) & -p_1 \exp(-\lambda_1) & -p_2 \exp(-\lambda_2) \end{bmatrix}$$

and **M** is the matrix with elements M_{ij} given by $M_{ij} = \frac{1}{n}\left(E(X^{i+j}) - E(X^i)E(X^j)\right)$. For *i,j=1, 2* these are the covariances of the sample moments (see Stuart and Ord, 1994). Also,

$$M_{33} = \frac{1}{n} g(0)(1 - g(0)),$$

$$M_{13} = M_{31} = -\frac{1}{n} g(0)\left[p_1\lambda_1 + (1 - p_1)\lambda_2\right]$$

$$M_{23} = M_{32} = -\frac{1}{n} g(0)\left[p_1({\lambda_1}^2 + \lambda_1) + (1 - p_1)({\lambda_2}^2 + \lambda_2)\right]$$

where $g(0) = p_1 \exp(-\lambda_1) + (1 - p_1)\exp(-\lambda_2)$, i.e. the probability of 0 for a 2-finite Poisson mixture. The above formulas can be derived using the appropriate formulas for the covariances of sample proportions given in Stuart and Ord (1994).

Example 3.2 (continued) Consider the data of Example 3.2, presented in Table 3.12. The large proportion of zeroes gives an indication that the method of zero frequency may be preferred as an estimating method. So, this methods was applied to these data along with the ML and the moment methods. The results are reported in Table 4.6.

From Table 4.6 one can see the estimates obtained by the three methods. The zero frequency estimates have standard errors which are very close to the standard errors of the ML estimates. The moment estimates have standard errors larger than those of both of the other methods. An explanation for this is the large right tail of the data which influences the moment estimation. From the practical point of view the interest for this data set lies mainly in the proportion of zero values which provides information about the proportion of 'good' clients. So, the method of zero frequency would be a natural choice for this data set, as it makes 'better' use of the information at the origin.

Table 4.6
The estimates of the parameters of a 2-finite Poisson mixture for the data of Table 3.12

Method	estimated parameters		
	p_1	$\ddot{e}_1$	$\ddot{e}_2$
Maximum Likelihood	0.768 (0.0075)	0.232 (0.0109)	6.156 (0.2889)
Moment	0.871 (0.0308)	0.469 (0.1561)	9.108 (1.1395)
Zero Frequency	0.826 (0.0075)	0.255 (0.0110)	7.893 (0.3109)

4.5.2 Comparison with Other Methods

Tables 4.7a-4.7c contain the asymptotic efficiency of the method of zero frequency relative to the ML method. It can be seen that for distributions with a high probability at 0 the efficiency is large. Comparing this method to the ordinary MM one can see that it is more efficient for distributions with a low mean. This makes this method an interesting alternative to the ML method when there are many counts at 0. Note also that the method succeeds in providing estimates which give an excellent fit for the 0 frequency. Note that as Harris (1991) pointed out, the estimated frequency at zero from the ML estimates is always greater than the observed frequency.

Figure 4.2 depicts the asymptotic efficiency of the zero frequency method for $\ddot{e}_1=1$ and various choices of the mixing proportion and the second parameter. The efficiency is high for components close to the positive axis, since such 2-finite Poisson distributions have a large proportion of zeroes. For cases with a low zero frequency, the method is not satisfactory.

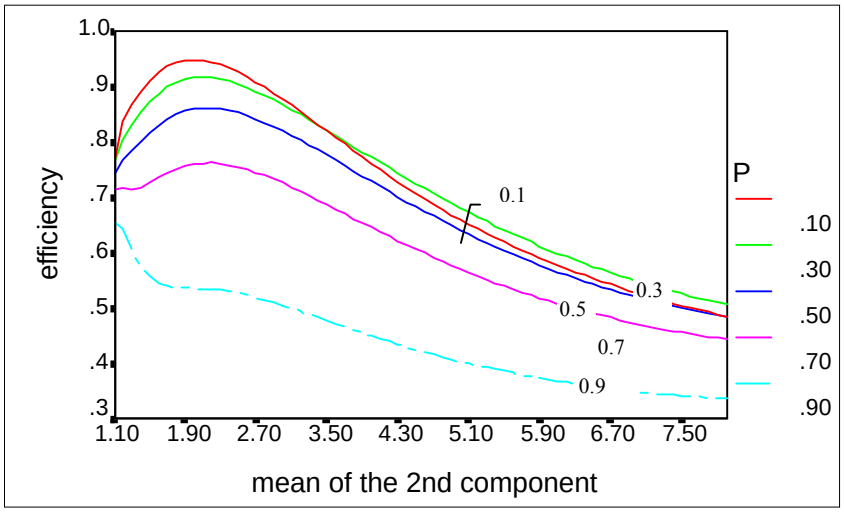


Figure 4.2 Asymptotic efficiency of the method of zero frequency for 2-finite Poisson mixtures with $\ddot{e}_1$=1

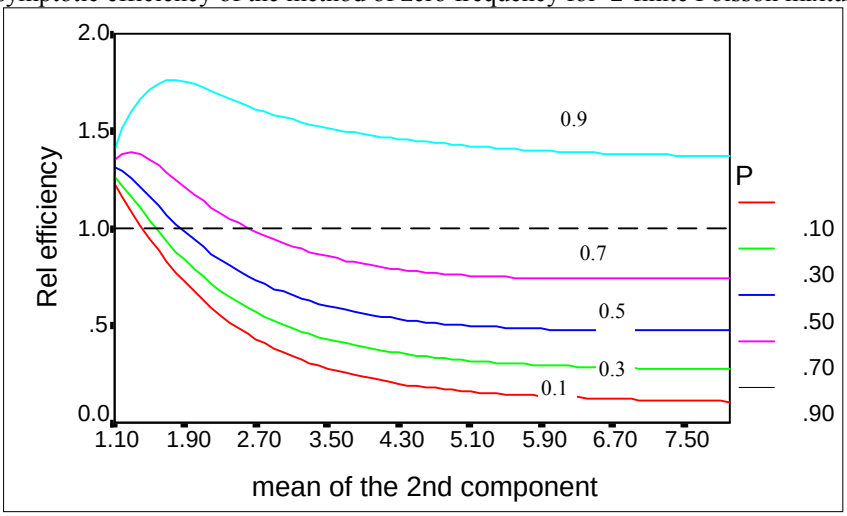


Figure 4.3 Relative efficiencies of the two methods. The efficiency was calculated as the ratio of the generalised variance of the moment method divided by that of the zero frequency method. Values smaller than 1 support the moment method.

Figure 4.3 compares the MM to the method of zero frequency by depicting the efficiency of the method of moments relative to that of the zero frequency method. Values lower than 1 favour the MM. When the mixing proportion is large, i.e. when the zero frequency is high, the zero frequency method is superior to the moment method. Perhaps, these two methods can be used complementary to fully exploit their behavioural features. Note that the ML method is preferable in all cases, apart from the situation where special attention is paid to the estimation of the zero frequency.

Table 4.7a
Asymptotic efficiency of the method of zero frequency relative to the ML method, for $\ddot{e}_1=1$

	$\ddot{e}_2$								
p_1	2	3	4	5	6	7	8	9	10
0.1	0.949	0.879	0.762	0.661	0.585	0.528	0.486	0.458	0.439
0.2	0.935	0.878	0.778	0.687	0.611	0.553	0.511	0.481	0.460
0.3	0.917	0.868	0.773	0.682	0.606	0.549	0.508	0.479	0.460
0.4	0.893	0.850	0.757	0.666	0.592	0.537	0.498	0.471	0.454
0.5	0.861	0.821	0.730	0.642	0.572	0.520	0.484	0.460	0.445
0.6	0.819	0.782	0.695	0.612	0.547	0.500	0.467	0.446	0.433
0.7	0.761	0.727	0.647	0.572	0.514	0.472	0.445	0.427	0.416
0.8	0.677	0.645	0.576	0.513	0.465	0.432	0.410	0.396	0.389
0.9	0.537	0.505	0.451	0.405	0.372	0.350	0.337	0.330	0.328

Table 4.7b
Asymptotic efficiency of the method of zero frequency relative to the ML method, for $\ddot{e}_1=2$

	$\ddot{e}_2$								
p_1	3	4	5	6	7	8	9	10	11
0.1	0.753	0.764	0.680	0.571	0.474	0.400	0.346	0.308	0.281
0.2	0.712	0.708	0.621	0.520	0.435	0.372	0.326	0.294	0.271
0.3	0.669	0.653	0.567	0.474	0.399	0.344	0.305	0.278	0.258
0.4	0.624	0.598	0.516	0.432	0.367	0.319	0.285	0.262	0.246
0.5	0.575	0.542	0.466	0.392	0.335	0.294	0.265	0.245	0.232
0.6	0.521	0.483	0.414	0.350	0.302	0.267	0.243	0.227	0.216
0.7	0.461	0.418	0.357	0.304	0.264	0.236	0.217	0.205	0.196
0.8	0.389	0.342	0.290	0.248	0.217	0.196	0.182	0.173	0.168
0.9	0.295	0.240	0.201	0.171	0.150	0.136	0.127	0.122	0.120

Table 4.7c
Asymptotic efficiency of the method of zero frequency relative to the ML method, for $\ddot{e}_1=3$

	$\ddot{e}_2$								
p_1	4	5	6	7	8	9	10	11	12
0.1	0.525	0.546	0.482	0.391	0.311	0.252	0.210	0.181	0.160
0.2	0.484	0.479	0.410	0.330	0.266	0.219	0.187	0.164	0.148
0.3	0.444	0.424	0.356	0.287	0.234	0.195	0.168	0.150	0.136
0.4	0.405	0.374	0.312	0.252	0.207	0.175	0.153	0.137	0.126
0.5	0.366	0.329	0.272	0.221	0.183	0.156	0.138	0.125	0.116
0.6	0.326	0.284	0.233	0.191	0.159	0.137	0.122	0.112	0.105
0.7	0.284	0.238	0.194	0.159	0.134	0.117	0.105	0.097	0.091
0.8	0.238	0.188	0.152	0.124	0.105	0.092	0.083	0.077	0.074
0.9	0.183	0.127	0.100	0.081	0.068	0.060	0.054	0.050	0.048

4.6 Bayesian Methods of Estimation for Finite Mixtures

Till now, we have only presented classical estimation methods. Bayesian methods have also been proposed for estimation of the parameters of finite mixture models.

Bayesian statistics treats the parameters to be estimated as random variables having their own distribution, known as the prior distribution. The estimation is based on the posterior distribution of these parameters given the data, calculated via the Bayes Theorem. This fact can cause a lot of difficulties, since the derivation of the posterior is not always straightforward, especially in the case of finite mixtures. The calculation of the posterior distribution is quite tedious in mixture models. This has always impeded the application of Bayesian methods until recently, when sophisticated sampling based approaches disentangled the problem. Much research has been made in the area of Bayesian estimation methods for mixture models using the fashionable Markov Chain Monte Carlo methods and a lot of interesting results have been obtained.

A first attempt for Bayesian estimation can be found in Rolph (1968). In order to construct a prior distribution for a parameter belonging to the interval (0,1) he makes use of the relationship between the moments of the mixed distribution to those of the mixing distribution. A similar idea is given in Meeden (1972).

To proceed in a fully Bayesian context one needs to specify the prior distributions of the parameters to be estimated and then to calculate the posterior distributions. According to Diebolt and Robert (1994), given a proper prior, a Bayesian approach to the problem always provides estimators which can be written explicitly for conjugate priors. However, standard Bayesian approaches, must be regarded as of purely academic interest and not of practical use. The reason is that the volume of computations involved is prohibitively large even for small sample sizes and increases exponentially with the sample size. A detailed evaluation of the posterior distribution requires too much computing time to be regarded as a plausible solution.

For obtaining the posterior distribution, all the possible partitions of the sample into components must be constructed. For example, for a 2-component model and a sample of size n, one can partition the entire sample in 2 components by

$\sum_{i=0}^{n}\binom{n}{i}$, different ways. Improper priors cannot be used since they will lead to inconsistent results.

In order to overcome this difficulty with a pure Bayesian approach, Smith and Makov (1978) discussed a quasi-Bayesian approach. Since their main interest was to identify the component from which the observation came, they only tried to estimate the mixing proportions; thus their procedure cannot be applied to the more general case where the parameters of the subpopulation distribution must be estimated too. The method requires the computation of the posterior density for each observation separately, in a successive manner. A disadvantage of this quasi-sequential method is that the results obtained depend on the order in which the observations are considered and thus its applicability can be limited to cases where there is a natural ordering of the observations.

Redner *et al.* (1987) described modal estimators for the parameters using conjugate priors for the component distributions. The equations for all the estimators can be written explicitly. However, solving these equations requires special numerical methods. Redner *et al.* (1987) proposed an iterative algorithm for solving them.

An advantage of this procedure is that the imposition of a prior leads to avoiding problems connected with the unbounded likelihood function. Moreover, the researcher can incorporate easily his own belief by choosing an appropriate prior. Problems arise if the posterior distribution is multimodal whence multiple modes exist. Recent research has demonstrated that this is the usual case for mixture models. The posterior distributions are usually multimodal due to the ambiguity about the number of components. Special effort on choosing the priors is needed to avoid multimodality.

Bernardo and Giron (1988) discussed in detail all the problems occurring when applying fully Bayesian methodologies identifying possible directions for progress.

Another approach to overcome the difficulty for the complete evaluation of all partitions is the use of contiguous partitioning described in Aitkin *et al.* (1996). They proposed a test for detecting the presence of a mixture based on posterior Bayes factors. To apply their method one needs to classify each observation. Contrary to the fact that complete enumeration of all possible partitions is impossible, even for moderate sample sizes, contiguous partitioning requires the mere determination of the

cut points which partition the entire sample to subsamples. By this, only n-1 different partitions are considered, making the calculations feasible.

Rajagobalan and Loganathan (1991) proposed another quasi-Bayesian method for estimating the mixing proportion. Since full Bayesian estimation is not computationally feasible for practical purposes they gave a compromise solution. For a given prior, taken by the authors to be a Dirichlet distribution, they calculated the posterior probabilities of each observation to belong to the i-th subpopulation and they subsequently used averages of these probabilities as estimates of the mixing proportions.

Crawford (1994) proposed the use of approximate methods for Bayesian estimation of finite mixtures using the Laplace method. The Laplace method was introduced in Bayesian statistics by Tierney and Kadane (1986) and Tierney *et al.* (1989). The Laplace method itself is a method for calculating the ratio of two integrals of a specific form. Usually, the posterior moments of a random variable are expressed as the ratios of two integrals and thus, the method is applicable. This method often leads to very good approximations of several posterior measures of interest avoiding the exact calculation which requires the evaluation of all possible partitions of our sample. The method works well when the posterior is similar to a Normal distribution.

Markov Chain Monte Carlo (MCMC) methods led to a phenomenal increase in the application of Bayesian methodology. The main gain is that the volume of required calculations is avoided by merely sampling from the target posterior distribution. This is feasible with a little effort.

The Data Augmentation algorithm (Tanner and Wong, 1987 , Wei and Tanner, 1990) is such a method. The Data Augmentation uses the missing data representation of finite mixture models and can be considered as a next step after the SEM algorithm described earlier.

Suppose that our full data representation Y contains an observed part and a missing part (or a part which can be considered as missing), i.e. $Y_i = (X_i, z_i)$ where X_i is the observable part of our data and the z_i is the missing part. The vector z_i has elements z_{ij}=1 if the i-th observation belongs to the j-th subpopulation and 0 otherwise. Thus, the 'missing data' are the subpopulation members of the observations

(see also the missing data representation used for constructing the EM algorithm in section 3.3.2). The algorithm proceeds with two steps:

At the first step the vectors z_i are simulated from a multinomial distribution, with n=1 and probabilities equal to the posterior probabilities

$$w_{ij} = \frac{p_j f(x_i|\lambda_j)}{f_P(x_i)}$$

i.e., the posterior probabilities that the i-th observation belongs to j-th group. This step is the same as the first step of the SEM algorithm.

The second step, instead of maximising the likelihood conditional on the memberships of the first step, simulates the parameters from the joint posterior distribution of the parameters given the data and the vectors z_i . The new parameter values are then the means of these posterior distributions. In other words, the M-step of the SEM algorithm is replaced by a step which proceeds in the usual Bayesian approach by estimating the parameters with the means of the posterior distributions. The missing data presentation is very similar to the one used in section 3.3.

The main problem of the Data Augmentation algorithm is the specification of the joint posterior distribution, which usually is not simple.

Recent methods in Bayesian analysis, overcome the problem of expressing the posterior distribution in a closed form (or even in an analytical form) by simply sampling from this distribution. The "Gibbs Sampler" is such a scheme, which can be applied to mixture problems as well. The only task is to determine all the marginal distributions wanted and then sample from them. Lavine and West (1992) demonstrated this idea in a problem of clustering which is very similar in nature to mixture problems.

Diebolt and Robert (1994) used the same idea in mixture problems. They showed the hierarchical nature of the problem and they applied such a procedure to a normal mixture. This technique seems to simplify the problems with a fully Bayesian approach. The conditional distributions involved can easily be determined since the Dirichlet distribution is the conjugate for the mixing proportions and we can use the appropriate conjugate for the parameters, too. For example the use of a Gamma distribution for the parameters of the Poisson distribution is plausible. A known application of such an estimation procedure can be found in Dellaportas *et al.* (1995) for finite mixtures of normal distributions.

Escobar and West (1995) proposed basing priors in Dirichlet processes. This approach enables us to simultaneously check for several models with different number of parameters.

The main difference between the Data Augmentation algorithm and the Gibbs sampler is that the former simulates form the joint posterior distribution while the latter simulates successively from the conditional distributions.

It is worth noting that all four methods (classical or Bayesian), namely the EM algorithm, the SEM algorithm, the Data Augmentation method and the Gibbs Sampler, make use of the 'missing data' representation of finite mixtures. They are based on two steps which lead either to closed form solutions or to the derivation of the solution via simulation. Table 4.8, summarises all the methods. A comparative description of the methods can be found in Tanner (1991).

Table 4.8
Algorithms for the estimation of the parameters of finite mixture models

algorithm	E-step	M-step
EM	closed form	closed form
SEM	simulated	closed form
Data Augmentation	simulated	simulated from the joint posterior
Gibbs Sampler	simulated	simulated from the conditionals

Note: The M-step for the Data Augmentation and the Gibbs Sampler algorithm is not a maximisation step, but a step which extracts the posterior density of the parameters.

Recently, Richardson and Green (1997) proposed a reversible jump Markov Chain Monte Carlo method for finite mixtures with an unknown number of components. They treated the number of components as an unknown parameter which has to be estimated from the data. This idea of treating the number k as an unknown parameter, has also been demonstrated in the Bayesian context by Binder (1978) in an application of Bayesian cluster analysis. The reversible jump MCMC is a flexible method which allows for jumping between models with different numbers of parameters.

Leonard *et al.* (1994) also proposed Bayesian estimation using an alternative Gibbs sampler known as Permutable Bayesian Marginalisation. In this paper, the mixing proportion was set equal to 1/k for k-finite mixtures.

Mengersen and Robert (1996) reparametrized the normal mixture model in such a way so that the prior distribution can be easier handled. Using Gibbs sampling they derived the posterior distributions and used them for testing the existence of a mixture.

Robert (1996) provides a thorough description of Bayesian methodologies for mixture models

4.7 Empirical Bayes

In Bayesian estimation, the parameter is not a constant number but a random variable with some distribution. Usually, this distribution is assigned to the parameter, either subjectively by the researcher on the basis of the researcher's prior knowledge and personal beliefs, or in a specified form which contributes to numerical simplicity and to the tractability of the problem. The non-informative prior distribution is a common choice representing the ignorance of the researcher concerning the behaviour of the parameter.

This specific aspect of the Bayesian approach has been criticised strongly. The choice of the prior is subjective and, hence, different researchers with the same data can give different answers simply because of different prior beliefs.

A mild and compromising approach is provided by the so-called empirical Bayes methods. According to them the prior distribution of the parameter is estimated from the data.

The connection with our problem is obvious: The purpose of the empirical Bayes methods is to estimate the mixing distribution, though the aim is different. In the empirical Bayes context, the mixing distribution is termed as the prior distribution and the mixed distribution is usually termed as the predictive distribution. One difficulty that may arise relates to the fact that the aim of empirical Bayes methods is not to determine the mixing distribution itself but rather to use it. As a result, using empirical Bayes methods may not lead to an explicit solution for the form of the mixing distribution as their estimation step is confounded with other steps (e.g. Maritz, 1969).

The duality of all the methods has to be stretched. Every method of estimating the mixing distribution can be used for empirical Bayes purposes (see, e.g., the books of Maritz and Lwin, 1989, Carlin and Lewis, 1996). On the other hand, several empirical Bayes methods lead to an estimate of the mixing distribution. For example, Laird (1982) proposed the use of the ML estimate of the mixing distribution as the prior distribution in an empirical Bayes application. Similarly, we may use another method of estimation as, for example, the MM used by Maritz and Lwin (1989). This approach utilises the estimation methods for mixture models in empirical Bayes

applications. This is also the case with mixture models which lend their methods to empirical Bayes problems. The opposite is also possible. For many empirical Bayes applications, each observation is related to its parameter. We may consider these estimates of the parameter to construct the mixing distribution. (see, e.g. , Efron and Morris, 1975, Lemon and Krutchkoff, 1969 and Robbins, 1964, 1983, among others).

Several empirical Bayes methods have been proposed for the Poisson distribution in a variety of articles (see, for example, Clevenson and Zidek, 1975, Gaver and O'Muircheartaigh, 1987, Maritz and Lwin, 1989, Walter and Hamedani, 1991, among others). No details will be given for these particular methods, since this is beyond the scope of this short review.

4.8 Miscellaneous Methods

We conclude our brief review of methods of estimation for finite mixtures by describing some miscellaneous methods, the applicability of which has not yet been demonstrated.

Zhang (1990) proposed the use of Fourier methods for estimating the mixing distribution. He used the well known inversion theorem for the characteristic function. For certain continuous and location families, e.g. the normal family , the Cauchy family etc., it can be shown that the characteristic function of the mixing distribution can be determined from the characteristic function of the mixed random variate. Hence, using the empirical characteristic function, one can apply the Fourier inversion theorem. The problem now is to choose a suitable kernel for reconstructing the mixing density. The author described a selection procedure as well as rates of convergence and lower bounds for the optimal rate of convergence. However he did not provide any examples to demonstrate the practical value of the method.

More recently, Zhang (1995) presented similar results for discrete exponential families, including the Poisson case. His approach is based on the well known property that the moments of the mixing distribution are related to the moments of the mixed distribution. Since the characteristic function can be related to the moments of a distribution, he introduced kernels in order to reconstruct the probability function by a Fourier inversion formula. For the Poisson case, he proposed to estimate the mixing distribution by:

$$\hat{g}_n(a) = \frac{e^a}{n}\sum_{i=0}^{n} K_n(X_i, a)\,.$$

Here, n is the sample size, X_i , i=1,..,n are the observations and $K_n(x,a)$ is a suitable function defined by

$$K_n(x,a) = \frac{1}{\pi}\int \cos(ta)t^x \cos(\frac{x\pi}{2})k^*\left(\frac{t}{c_n}\right)dt$$

where k^* (t) is a function such as $k^*(-t)=k^*(t)$, $\forall t \in [0,1]$ t, and $k^*(t)=0$, for $|t|>1$, and c_n is a constant whose nature is similar to that of the bandwidth in kernel density estimation. For the choice of $k^*(t)$ one may proceed in a manner similar to that employed in kernel density estimation.

The author also studied the rate of convergence of his estimator. Obviously, this estimator is not easy to apply because of the complexity of the involved integral in its expression. Variants of the above estimator can be found in Loh and Zhang (1996, 1997). Goutis (1997) developed another kernel based estimator which is relatively simpler than those described above.

Other miscellaneous methods include those proposed by Preston (1971) who used piece-wise polynomial arcs to construct an estimate of the mixing distribution, by Walter (1985) who used orthogonal polynomials for estimating the mixing distribution of Poisson mixtures and by Hengartner (1997) who used an estimator based on kernel estimates via orthonormal polynomials .

Chapter 5

Minimum Hellinger Distance Based Inference for Finite Poisson Mixtures

5.1 Minimum Distance Estimation

5.1.1 General Introduction

It is known, for mixture models that samples of small size and components close together do not provide good estimates either through the method of moments or through the ML method. In order to overcome this, a lot of research has been carried out for alternative estimation methods, particularly for minimum-distance methods. In fact, one can use some kind of distance between the observed data and the expected data and then try to solve the problem by minimising this distance. As will be seen later, both the moment method and the ML method can be regarded as minimum distance methods. In this sense, the estimation problem can be reduced to one of choosing the distance function with some optimality criteria.

Minimum distance estimation is used in parametric inference when the model is suspected to be inexact and the existence of some observations far from the main body of data may cause a lot of trouble in the estimation. Unfortunately, not much work has been done for examining robustness in mixture model analysis. Such examples are the works of Gray (1994) and Gustafson (1996). The former examined the bias when the component distributions do not belong to the assumed family of distributions, while the latter treated the case where the mixing distribution assumed is incorrect. Methods which are robust with respect to such departures from the model are very useful. Depending on the distance function considered, we can cope with these outliers when the classical ML method fails to do so.

A distance $\delta(f,g)$ is defined between the functions f and g, which measures, in some sense, how far the function f from the function g is. Usually, f and g are probability functions or probability density functions, but for some methods distribution functions or moment generating functions can be used as well. Very often, weighted versions of the distances are used, weighting with respect to some function, say $w(x)$. In this manner the interest in some points of the real line is increased by putting more weight to certain neighbourhoods of x.

According to Parr and Schuccany (1982), one can distinguish between two main categories of distance functions:

- those of an integral type, namely those which measure over the real line the discrepancy between the two functions, or some weighted versions of the functions and
- those based upon "sup-type" discrepancies, namely those which determine the maximum discrepancy over the real line.

For example, the well known chi-square distance is of the former type, while the Kolmogorov distance is of the latter.

The first category possesses some useful properties like asymptotic normality under suitable conditions, while for the second category, such useful properties are usually hard to be proven and usually the obtained estimators are not even asymptotically normally distributed.

Several distance measures have been proposed. A small collection can be found in Titterington *et al*. (1985, pp 116). In fact, one can use a lot of different functional forms for both the distance function and the weight function. Moreover, one can either use the distribution functions or the probability functions depending on the nature of the data under investigation. For example, for count data the use of the empirical relative frequencies as an estimate of the probability function is straightforward. For continuous data, however, this is not so since one needs to place observations in classes with arbitrarily chosen bounds or to estimate the density using some method like the kernel density estimation method.

From the above, it is obvious that the number of different choices of distance functions is very large. This complicates the estimation problem particularly since comparative results for all these functions are not known and one is not able to

comparatively judge their performance. Lindsay (1994) tried to compare some of them on the basis of their relative efficiency.

In the sequel, a variety of distances will be reviewed. It is useful to note that both the MM and the ML method can be considered to be minimum distance methods. Kemp (1986) showed that many of the known procedures for estimation can be regarded as methods of weighted discrepancies between the observed and the expected data. According to Kemp (1986) the methods can be separated in three main categories: those with constant weights (not depending on parameters or observed frequencies) like the moment method, those with weights depending on the parameters, like the ML method, and those whose weights depend on both the parameters and the observed frequencies like the minimum chi-square method.

The ML method uses the so called Kullback-Leibler distance. The Kullback-Leibler distance is defined by

$$\begin{aligned}\delta_{KL}(f,g) &= \sum_x g(x)\ln\frac{g(x)}{f(x)} = \\ &= \sum_x g(x)\ln(g(x)) - \sum_x g(x)\ln(f(x))\end{aligned} \quad . \qquad (5.1)$$

If $g(x)$ denotes the observed relative frequency of the value x, the minimisation of this distance is equivalent to the maximisation of the second term in (5.1) which is proportional to the well known loglikelihood of a sample. Therefore, the minimum distance estimates using the Kullback-Leibler distance are the same as the ML estimates.

As far as the method of moments is concerned we can define the distance

$$\begin{aligned}\delta_{MM}(f,g) &= \sum_{t=1}^{k}\int_x \big(f(x)-g(x)\big)w(x,t)dx = \\ &= \sum_{t=1}^{k}\int_x f(x)w(x,t)dx - \sum_{t=1}^{k}\int_x g(x)w(x,t)dx = \\ &= \sum_{t=1}^{k} E_f\big(w(x,t)\big) - E_g\big(w(x,t)\big)\end{aligned} \qquad (5.2)$$

where $f(x)$ and $g(x)$ are density functions and $w(x,t)$ is a weight function. It is easy to show that if g is the empirical density and f some density, the minimisation of the distance is equivalent to equating the expected values of the functions $w(x,t)$.

Taking $w(x,t) = x^t$, for t=1,2,...,k , leads to the moment method using the first k moments.

Hall (1981) showed the above result. Generally speaking, the choice of the weight function is crucial in estimation with distance measures. This function must not give important weight far from the main body of data. A weight function of the form $w(x,t) = x^t$ clearly gives weights that increase with x and this is the reason for the inadequacy of the method of moments, especially when a large number of moments is used.

From the above discussion it becomes obvious that all the known estimation methods can be regarded as minimum-distance methods. Among them are, the ML method and the method of moments which have played an important role in statistical estimation and have been already presented in chapters 3 and 4 respectively.

5.1.2 Minimum Distance Methods Applied to Finite Mixture Estimation

Recall that a general point about the estimation of the mixing distribution is the restriction to estimate it by a step-function even though the true mixing distribution is continuous (Laird, 1978). All the methods of this section require the number of support points to be known. In the case where the number of support points is not known, the methods can be applied for several values of the support size and then the one satisfying some optimality criteria can be chosen (see, e.g., Chen and Kaldbfleisch, 1996). Semiparametric minimum Hellinger estimation for finite Poisson mixtures is discussed later in this chapter.

The interest of most of the cited papers is focused in estimating the mixing proportions assuming that the component distributions are known. This approach eases much of the computational difficulty. However, the methods can be extended to include the estimation of the parameters of the components. In this case, however, the computational effort involved is much greater.

In the sequel, *F(x)* denotes the assumed distribution function, while *G(x)* denotes the empirical distribution function calculated from the data. The probability function and the empirical relative frequency are denoted by *f* and *g* respectively.

Boes (1966, 1967) tried to estimate the mixing proportion *è* for the case of a 2-finite mixture with known components, namely a mixture of the form

$\theta F_1(x) + (1-\theta)F_2(x)$, where F_1 and F_2 are known distribution functions. The only parameter to be estimated is *è*. He proposed the obvious estimator

$$\hat{\theta}_x = \frac{G(x) - F_2(x)}{F_1(x) - F_2(x)} \quad . \tag{5.3}$$

Obviously, from (5.3) one can find an estimator for *è* for every value of x. Boes proposed to use an averaged estimator of the form $\int_x \hat{\theta}_x w(x)dx$, where $w(x)$ is some weight function. He discussed and proposed a specific form for $w(x)$. Some asymptotic properties of the estimator were derived and its behaviour for small samples was examined. It was found that the estimator satisfies some optimality criteria such as the Cramer-Rao bound for its variance. Ahmad *et al*. (1983) proposed another related estimator based on the derivation of (5.3) for selected quantiles. The proposed estimator was an average of these quantile-based estimators. Later, Van Houwelingen and De Vries (1987) developed a shrinkage estimator based on (5.3) which is minimax.

Choi and Bulgren (1968) and Choi (1969), considered minimising the averaged L_2-norm. Note that averaged L_2 -norm distance W(F,G) between two distribution functions F and G is defined as

$$W(F,G) = \int \left(F(x) - G(x)\right)\, dF(x)\,.$$

By some authors, this distance is termed as the Wolfovitz distance while Titterington *et al*. (1985) use this term for another type of distance.

For estimation purposes, one needs to use the discrete analogue of this distance, namely the function

$$S_n(P) = \frac{1}{n}\sum_{i=1}^{n}\left(F_P(x_{(i)}) - \frac{i}{n}\right)^2 \quad , \tag{5.4}$$

where $x_{(i)}$ is the i-th order statistic from the sample. The subscript P refers to the mixing distribution that leads to the mixed distribution F_P. In the case where only the mixing proportions are estimated it is assumed that the support points of the mixing distribution are known.

Estimating the mixing distribution P is equivalent to determining a function $\hat{P}$ such that the quantity $S_n(P)$ is minimised. Unfortunately, minimising (5.4)

requires solving a linear optimisation problem. The bigger the number of parameters the more difficult to solve the problem. Choi and Bulgren (1968) estimated the mixing proportions only and they provided a few simulations concerning the behaviour of their estimators. They concluded that their estimator seemed to work reasonably well for moderate sample sizes. They also investigated the asymptotic properties of these estimators and showed that, under some conditions, it is consistent. Later, Henna (1983) showed that the conditions considered by Choi and Bulgren (1968) can be relaxed without affecting the consistency of the estimator. Henna (1985) proposed a sequential minimisation of this distance in order to determine the number of components in the mixture. Note that for both proofs F_P is required to be continuous. The consistency of the estimator for discrete F_P is not known. The same distance was used by Robbins (1964) for an empirical Bayes problem.

Bartlett and McDonald (1968) proposed a "least square" distance of the form

$$\delta_{BM}(F,G) = \int (dG - dF)^2 / dH \ ,$$

where H is an increasing function of x. They showed that with a suitable choice of the weighting function the variance of the estimates can be decreased, and they proposed several such weights.

MacDonald (1971) criticised the estimator of Choi and Bulgren which minimises (5.4) and proposed another distance which has a smaller bias. He also compared the two distances with a few simulations. His idea is based on using the Cramer Von-Misses distance $CVM(F,G)$, defined as:

$$CVM(F,G) = \int \left(F(x) - G(x)\right)^2 dG(x) \qquad . \tag{5.5}$$

The sampling counterpart of this distance is

$$S_n(G) = \frac{1}{12n} \sum_{i=1}^{n} \left(F(x_{(i)}) - \frac{i - 0.5}{n} \right)^2 \qquad . \tag{5.6}$$

The similarity to the Wolfovitz distance is obvious. The Cramer-Von Mises distance is based on the difference between the cumulative distribution functions averaged with respect to *G(x).* The term 0.5 is used to correct for approximating a discrete function by a continuous one.

This distance was also examined by Woodward *et al*. (1984) who made a simulation comparison of it to the ML estimator in the case of normal mixtures. They

found that Minimum Distance estimation based on the Cramer-Von Misses distance given in (5.6) leads to better estimators under symmetric departures from normality thus concluding that they are more robust than the ML estimators.

Clarke (1989) used this distance for estimating only the mixing proportion. He showed that the estimator can be written in a closed, though complicated, form and that the method is robust when the actual underlying models are mixtures of heavy tailed densities, like the exponential or the student-t distributions.

A quite similar method is proposed by Deely and Kruse (1968). They tried to minimise the sup norm $SN(F,G) = \max_x |F(x) - G(x)|$. It can be shown that the problem is a linear programming problem with a great number of constraints. Hence, the computational effort is tremendous. Because of the complexity of this approach, it has not been used in practice. The authors also showed that the estimated mixing distribution tends to the true mixing distribution with probability one when the sample size is large. A few years later Chandra (1977) extended their method. Chen (1995) showed that these estimators can achieve the best possible rate of convergence to the true mixing distribution which is $n^{-1/4}$.

A similar approach is given by Phillips (1990). He proposed to use the Chebyshev norm instead of the sup norm. He showed that in this case the linear programming problem has a smaller number of constraints and therefore it is easier to be solved. He also showed that the estimated mixing density tends, with probability 1, to the true one. He gave some illustrating examples applying his method.

Blum and Susarla (1977) proposed a method which involves a system of inequalities to be satisfied simultaneously. Their method tries to minimise a distance similar to the Kolmogorov distance. The authors pointed out that the problem was a linear programming problem. The applicability of their method is limited since the computational effort is large even with high-speed computer devices.

Fryer and Robertson (1972) proposed the well known minimum chi-squared distance for estimating the mixing distribution. Comparing this method with the moment method they found that the procedure is similar in efficiency. The problem with such a method is that the data must be grouped. With discrete data this is straightforward, but with continuous data it may lead to bias depending on the different categorisation.

Albrecht (1982), discussed several methods for estimating the parameters of finite Poisson mixtures. Among them he proposed a quasi-minimum chi-square method. He showed that the numerical complexity of this method is not prohibitive as it could be for the full minimum chi-square technique. The method may be called quasi-minimum chi-square because it expands the chi-square distance in a Taylor series and uses only the first two terms. The inaccuracy of the method is very small since terms larger than the second may be regarded as negligible. He also provided us with a hint about the numerical solution of this problem.

Hall and Titterington (1984) suggested to combine different estimators in order to improve the efficiency of the estimate. The idea is that in this case the variance of the estimator is a continuous function depending on the unknown components. So, by choosing the way of categorising the data one can improve the estimate. Note that for count data the method is not appropriate.

Edelman (1988) and Clarke and Heathcote (1994) proposed the use of the integral L_2 distance, namely

$$L_2(G,F) = \int \left(G(x) - F(x)\right)^2 dx \quad .$$

Eddelman (1988) assumed a n-finite mixture with mixing proportions equal to $1/n$ whose only the component distribution parameters had to be estimated. Clarke and Heathcote (1994) proposed a usual k-finite model. Both papers treated the normal case. However, the motivation for this approach was different. Eddelman proposed it in an empirical Bayes context, while Clarke and Heathcote for robust estimation of the mixing distribution. They showed that the estimating equations satisfy the conditions required so as to yield robust estimates.

Cutler and Cordero-Brana (1996) proposed the use of the Hellinger distance to produce minimum Hellinger distance estimators for finite normal mixtures. They showed the asymptotic normality and the robustness of such estimators in case of finite normal mixtures and they gave an algorithm to facilitate the estimation.

Titterington (1983) pointed out that by using distances of the square type it is possible to obtain explicit expressions for the estimates of the mixing proportions when the latter are the only parameters to be estimated. With other methods or when the parameters of the component distributions have to be estimated as well, they proposed the classical iterative method of Newton-Raphson or the Scoring Method.

These methods are good devices if good initial values are known, otherwise they may not converge at all.

Quandt and Ramsey (1978) used a modified procedure. They proposed to use the sum of the squared differences between the empirical moment generating function $H(t)$ and the true moment generating function $h(t;\theta)$, evaluated at several distinct points t_i, *i=1 , ..., k*. The empirical moment generating function $H(t)$ is given by

$$H(t) = \frac{1}{n}\sum_{i=1}^{n} \exp(tX_i) \ ,$$

where n is the sample size and X_i , *i=1, ..., n* are the n observations. So, the proposed method minimises the distance

$$M = \sum_{i=1}^{k} \left(H(t) - h(t,\theta)\right)^2$$

with respect to the vector of parameters **è**.

As the authors showed, this procedure is similar to a moment method but with weights varying for each moment. Moments of higher order are given less weight. The main difficulty with this procedure is the computational complexity for minimising this function.

Applying the procedure to normal mixtures they showed that the behaviour for small sample size is satisfactory. Later Kumar *et al*. (1979) proposed the use of the empirical characteristic function. The reason is that the characteristic function is not as smooth as the moment generating function, and thus it is more appropriate to find departures from the assumed distribution. This idea was further pursued by Bryant and Paulson (1983) for the estimation of the mixing proportion only. However, the use of the characteristic function may increase the computational effort required, and its sensitivity may cause difficulties in a number of instances as for example, locating local minima which are not global etc. An improved version of the method of moment generating function can be found in Schmidt (1982). He extended the idea of Quandt and Ramsey (1978) by considering the minimisation of a generalised sum of squares rather than an ordinary sum of squares. He did so because the terms in the simple sum of squares are correlated and the resulting estimators will not be efficient in the same manner in which the least-squares estimates are not efficient in regression models with correlated errors. He also examined in detail issues concerning the number of

different points where the moment generating function (or the characteristic function) must be evaluated.

In the case of Poisson (or more generally discrete) mixtures it is plausible to use the empirical probability function instead of the moment generating function or the characteristic function. The probability generating function has a behaviour similar to that of the moment generating function while, for discrete cases, it provides a very useful insight. It can be shown that minimising the sum of squared differences between the empirical and the true probability generating functions is equivalent to minimising a weighted function between the observed and the theoretical frequencies. The problem of obtaining estimators based on minimising a distance between the true probability generating function and its sample counterpart for finite Poisson mixture models, remains open.

For a special family of distributions Lindsay (1986) proposed another method in order to obtain an estimate of the mixing distribution. His method requires that $f(x|\grave{e})$ belongs to the exponential family. He restricted the mixing distribution to belong to such a family, so that the mixed distribution be a member of the two-parameter exponential family. Using this fact he estimated the parameters of the mixture using a least square method for a certain function of the resulting distribution. This method is semiparametric in the sense that although it is assumed that the distribution belongs to a certain family, one avoids giving a specific form for it. By this, the estimated distribution can be flexible enough. The methods of estimation of such models do not require the full knowledge of the probability function of the resulting distribution.

It is very interesting that of all the above distances very few have been considered for Poisson mixtures, and none has been applied to Poisson mixtures. All of them have been considered for normal mixtures.

A natural question arises at this point as to what the preferable distance is. No clear answer exists. However some guidelines for selecting a distance can be given.

An important issue is the type of the data. If the data are continuous, selecting distances which use the probability function creates a major difficulty as it is hard to estimate the probability density function of the dataset. Hence in this case, distances which use the cumulative density function are more appropriate. Otherwise, sophisticated methods are needed to estimate the probability density function.

On the contrary, for discrete data, the use of the observed frequencies, to estimate of the probability function, makes the use of distances based on the probability function appealing. The main reasons for that are:

- The extension to multivariate cases is easier
- The treatment of the discrete data case is more natural.

Note that, for some models, the assumed categories have no ordering. So, the distribution function is not well defined, as for example in the case of latent class models (Everitt, 1984a).

In concluding this review section, we provide in Table 5.1 a summary of the majority of distances used and described above. The names of the distances are those used by the authors.

Table 5.1

Summary of distance measures considered for estimation of mixture models

Distance	Name	References
$\int \left(F(x)-G(x)\right)^2 dF(x)$	averaged L_2-norm	Choi and Bulgren (1968) Choi (1969) Henna (1983, 1985) Robbins (1964)
$\int \left(F(x)-G(x)\right)^2 dG(x)$	Cramer-Von Mises	McDonald (1971) Woodward *et al*. (1984) Clarke (1989)
$\int \left(F(x)-G(x)\right)^2 dx$	Squared Distance	Edelman (1988) Clarke, Heathcote (1994)
$\int \left(f(x)-g(x)\right)^2 w(x)dx$	Weighted L_2-norm	Bartlett, Mac Donald (1968)
$\sup\lvert F(x)-G(x)\rvert$	Kolmogorov	Deely, Kruse (1968) Chandra (1977)
	Chebyshev	Phillips (1990)
$\sum_x \left(\frac{\left(g(x)-f(x)\right)^2}{f(x)}\right)$	Chi-Squared	Fryer, Robertson (1972) Albrecht (1982)
$\int \left[\sqrt{f(x)}-\sqrt{g(x)}\right]^2 dx$	Hellinger	Cutler, Cordero-Brana (1995) Eslinger *et al*. (1995) Karlis and Xekalaki (1998b)
$\sum_i \left(h(t_i)-H(t_i)\right)^2$	sum of squares between moment generating functions	Quandt, Ramsey (1978) Schmidt (1982)
$\sum_i \left(\phi(t_i)-\Phi(t_i)\right)^2$	sum of squares between characteristic functions	Kumar *et al*. (1979) Bryant, Paulson (1983)

Note: For all the entries of the table f(x) is the assumed probability density function, g(x) are the observed relative frequencies, F(x) is the assumed cumulative function, G(x) are the observed cumulative relative frequencies, w(x) is any weight function, h(t) and H(t) are the assumed and the empirical moment generating functions respectively and ö(t) and Ö(t) are the assumed and the empirical characteristic functions respectively.

5.2 Robustness in finite mixtures

Estimation of the parameters of a hypothesised model is a rather difficult task in the sense that there is not any globally accepted criterion for selecting the best estimation procedure. In general, every estimation method is superior in some aspect, but it is possibly inferior in some other. It has been customised in Statistics to prefer estimators which are efficient, namely estimators with standard errors as small as possible. However in many situations when the hypothesised model may be incorrect the notion of robustness is as crucial as that of the efficiency. So, two rather different criteria have to be used in order to describe the performance of the estimators, and these two criteria are just some members of a probably long list of criteria.

In the previous section we reviewed in details several distance methods considered for general finite mixtures. The reason is that minimum distance estimation methods are appealing in parametric inference, especially in cases where the model is suspected to be inexact. The robustness of such methods has made them viable alternatives to the widely used ML method. However, some problems remain unsolved. For example, the wide variety of different possible distances makes difficult the choice of the distance to be minimised. Problems arise in selecting both the functional form of the distance and the functional from the data which has to be used. Moreover, the minimisation for several distances is rather difficult for many models, while, at the same time, the ML method is easily available at a low cost and with a little effort. Every distance involved in the estimation procedure can measure adequately some specific departure among the data counts and their expected frequencies under the assumed models. So, depending on the distance considered, the method is optimal in some aspect and thus the usefulness of every distance depends on the aspect which is more important. For example, some distances are sensitive to the presence of outliers, others to long tailed distributions. In general, each distance has its pros and cons.

In trying to strike a balance between efficiency and robustness one seeks a method that combines both of these properties. Many robust estimators achieve robustness at some cost in first order efficiency. So, a trade off between these two issues is necessary. Minimum Hellinger Distance (hereafter MHD) estimation

method is a possible candidate as it combines both of these aspects. Lindsay (1994) shows that ML estimators and MHD estimators are members of a larger class of efficient estimators with various robustness and second order efficiency properties.

In the sequel, the MHD method for finite Poisson mixtures is treated. The results are not restricted to the simple estimation for finite mixtures, but MHD based inferences will also be proposed. The purpose is to exhibit the interesting properties of the MHD method in comparison to the widely used ML method and to propose the extensive use of the MHD based inference on several aspects of finite mixtures, such as hypothesis testing and semiparametric estimation.

The presentation of the MHD method is organised as follows. A simple motivating example is given to demonstrate the imperative need for robust methods. Later in this chapter, it is shown that ML based methods can lead to inconsistent results when just one extreme observation (outlier) has been added in our sample. MHD estimators for finite Poisson mixtures are derived and their properties are examined. An extensive simulation comparison of the MHD method to the ML method is presented. This comparison covers both the aspects of efficiency and that of robustness. An application to real data illustrates how these two methods can provide different results for the same dataset. MHD based inferences for finite mixtures are also introduced including semiparametric estimation using the MHD method. This new method generalises some of the results of the section 3.7. Finally, a method for graphically determining the number of components is given. This leads to a very interesting graphical device for checking for the Poisson distribution for a dataset. The next chapter is devoted to developing a test statistic based on the MHD and to showing that it is both efficient and robust compared to the well known likelihood ratio test statistic.

Example 5.1 Before getting into the details about the MHD method let us consider the following artificial example that motivated our research. A sample of size 25 was drawn from a 2-finite Poisson mixture with true parameters $p_1=0.5$, $\ddot{e}_1=1$ and $\ddot{e}_2=3$. Table 5.2 contains the observed frequencies.

Table 5.2
The observed frequencies for a simulated sample of size n=25, from a 2-finite Poisson mixture with p_1=0.5, $\ddot{e}_1$=1 and $\ddot{e}_2$=3

x	0	1	2	3	4	5	6
frequency	8	4	5	1	3	2	2

In order to see how an outlier can destroy the results taken from this sample, we contaminated our data by adding a new observation far away from the bulk of the data. In particular, we added a 26th observation with value x_{26}=12. Table 5.3 contains the parameter estimates for the uncontaminated (original) and the contaminated model by both the ML method and the MHD method of estimation.

Table 5.3
ML estimates and MHD estimates for the uncontaminated (1) and the contaminated (2) data.

	model	p_1	$\ddot{e}_1$	$\ddot{e}_2$
ML method	uncontaminated	0.470	0.480	3.425
	contaminated	0.591	0.768	4.813
MHD method	uncontaminated	0.409	0.354	2.992
	contaminated	0.399	0.368	3.043

We can see that the new observation (which may be considered as an outlier) influenced the ML estimates very much, while its influence on the MHD estimates is almost negligible. In some sense, the MHD method ″ignored″ this observation ″detecting″ that it was an outlier. This example gives an indication that MHD estimates may work better in situations where there are outliers, because they can detect if a spurious observation belongs to the hypothesised model (a 2-finite Poisson mixture in our case) or it is an outlier due to misrecording or to misspecified model.

It is interesting to see that if a 3-component Poisson distribution is considered, the ML estimates will be $p_1 = 0.4211$, $p_2 = 0.5363$, $p_3 = 0.0426$, $\ddot{e}_1 = 0.4062$, $\ddot{e}_2 = 3.3031$ and $\ddot{e}_3$=11.2976. The third component has led to a mixing proportion of 0.0426 and a parameter estimate of 11.2976. The contamination was effected by one of the 26 observations (x=12) representing almost 4% of our data. In other words, the outlier observation is regarded as an additional component of the model, while the other two components are very close to the values obtained before the contamination. This fact has lead researchers to the strategy of fitting one more component for the

outlier when using mixture methods to detect outliers (e.g. Aitkin and Wilson 1980). An example for the case of count data can be found in Harris and Basu (1994), where the contamination is considered as one more component. From the above results it becomes obvious that the MHD method can be a robust alternative to the ML method when the presence of outliers may cause problems to the meaningful application of the ML method.

In this illustrative example the sample size was small. In fact the behaviour is similar for larger sample sizes as demonstrated in the sequel. What is surprising is that a single observation which is inconsistent with the model suffices to drive the likelihood based estimates far from the true values, while the MHD estimates remain relatively stable. This useful robustness property makes the MHD method an interesting alternative to the ML method.

5.3 Minimum Hellinger Distance Estimation for Finite Poisson Mixtures

MHD estimation was introduced for the first time by Matusita (1954) but it was rather ignored until Beran (1977) examined in depth the case of MHD estimation for parametric models. The main reason is that, since it is a distance based on probability density functions, it is not easily applicable for continuous models when a smooth estimate of the probability density function has to be obtained via kernel density estimation. This fact makes the method computationally demanding for continuous models. Later, Eslinger and Woodward (1991) examined the application of the MHD method for the normal distribution, Tamura and Boss (1986) for a multivariate normal case, and, recently, Woodward *et al*. (1996) and Cutler and Cordero-Brana (1996) for finite normal mixtures. In all these papers a kernel density estimate of the probability density function was used. Note that Basu and Lindsay (1994) proposed a revised approach for which both the data and the model are smoothed with the same kernel.

For discrete models, Simpson (1987) described the MHD method and Hellinger distance based tests (see also Simpson , 1989). Lindsay (1994) examined minimum distance methods for discrete distributions in general, while Harris and Basu (1994) and Park *et al*. (1995) extended the use of the MHD method by

considering it as a penalised ML method (in the first paper) and by using combined distances (in the second paper). Robust minimum distance procedures are described in Basu *et al.* (1997). Related material treating general minimum distance estimation can be found in Read and Cressie (1988), where the Hellinger distance is a special case, and in Basu and Lindsay (1992) where computational issues are covered.

Suppose that $d(x)$ is the observed proportion of the value x from a sample of size n and $f_\theta(x)$ is the probability under the assumed model that the random variable X takes the value x, where **è** denotes the vector of parameters of interest. The MHD estimates for discrete data can be derived as the vector $\mathbf{è_{min}}$ which minimises the Hellinger Distance *D* given by

$$D(d, f_\theta) = \sum_{x=0}^{\infty} \left[\sqrt{d(x)} - \sqrt{f_\theta(x)} \right]^2 \qquad (5.7a)$$

$$= 2 - 2 \sum_{x=0}^{\infty} \sqrt{d(x) f_\theta(x)} \quad . \qquad (5.7b).$$

If $f_\theta(x)$ is assumed to be a k-finite Poisson mixture, its probability function is given by (3.2) and the vector of parameters **è** is given as **è** = $\left(p_1, p_2, \ldots p_{k-1}, \lambda_1, \lambda_2, \ldots, \lambda_k\right)$. Note that the vector **è** can be considered as defining a finite step distribution and this representation will be used in a next section. For obtaining MHD estimates for the parameters of the k-finite Poisson distribution the distance given in (5.7) has to be minimised. The natural way to do so is to equate the partial derivatives with respect to the parameters to 0 and to solve the resulting system of non-linear equations.

For each parameter $è_i$, i=1, 2, . . . , 2k-1 we have an equation of the form

$$\sum_{x=0}^{\infty} \sqrt{\frac{d(x)}{f_\theta(x)}} \frac{\partial f_\theta(x)}{\partial \theta_i} = 0 \ , \qquad \text{for } i=1, 2, \ldots, 2k-1 \ .$$

So, the system of estimating equations is

$$\sum_{x=0}^{\infty} \sqrt{\frac{d(x)}{f_\theta(x)}} \left(f(x, \lambda_j) - f(x, \lambda_k) \right) = 0 \ , \qquad j=1, 2, \ldots, k-1, \qquad (5.8)$$

for the partial derivatives with respect to the mixing proportions and

$$\sum_{x=0}^{\infty}\sqrt{\frac{d(x)}{f_\theta(x)}}p_j\Big(f(x-1,\lambda_j)-f(x,\lambda_j)\Big)=0\ , \qquad j=1, 2, \ldots, k \tag{5.9}$$

for the component parameters λ_j, where $f(x,\lambda)=\exp(-\lambda)\lambda^x/x!$, namely the probability function of a Poisson distribution with parameter ë.

An analytical solution of the system of equations (5.8) and (5.9) is not feasible. Numerical methods are required to solve it.

5.4 Properties of the Minimum Hellinger Distance Estimators

In the sequel some properties of the estimators are examined. All of these properties are consequences of the theorems given by Simpson (1987) for MHD estimators of discrete distributions. In order to assure the identifiability of the parameters, the $ë_i$'s need to be assumed to be in ascending order as Teicher (1963) showed. This identifiability assumption is necessary for the application of Simpson's theorems.

Theorem 5.1 (Simpson, 1987). Suppose that $f_\theta(x)$ is continuous in *è* for each *x*. Then for each distribution function *F* which is not singular,
a) the MHD estimate *T(F)* exists, and
b) if *T(F)* is unique then $D(d,f_\theta)\to 0$ implies that *T(F)* is consistent.

In real applications *F* is the empirical distribution function and the functional *T(F)* is the MHD estimate. Applying the results of Theorem 5.1 , the MHD estimators for k-finite Poisson mixtures exist since the k-finite Poisson mixtures are continuous in their parameters for each x. Their consistency is also a consequence of Theorem 5.1. The class of finite mixture distributions is identifiable, and thus the estimates are consistent and asymptotically unbiased. Simpson (1987) imposed some smoothness conditions on the derivatives of $f_\theta(x)$ to prove the asymptotic normality of the estimators. The probability function of the Poisson distribution satisfies these

conditions. The conditions are also satisfied by the k-finite Poisson mixtures since the derivatives of the probability function of a k-finite Poisson mixture with respect to the parameters are also linear functions of the probability functions of the component Poisson distributions. In fact, this holds for mixtures of any distribution for which Simpson's (1987) conditions are satisfied. From the above argument and Simpson's Theorem 2, the asymptotic normality of the estimators can be established. So, the MHD estimator follows asymptotically a $MN(\mathbf{\grave{e}}, \mathbf{V})$ distribution, i.e. a multivariate normal distribution with mean vector $\mathbf{\grave{e}}$ and variance-covariance matrix $\mathbf{V}$. The variance covariance matrix $\mathbf{V}$ is calculated as $\mathbf{V}=\mathbf{H}^{-1}\mathbf{I}(\theta)\mathbf{H}^{-1}$, where $\mathbf{H}$ is the matrix with its ij-th element equal to

$$\mathbf{H_{ij}} = \frac{\partial^2 D(d, f_\theta)}{\partial\theta_i \partial\theta_j} = -\frac{1}{4}\sum_{x=0}^{m}\sqrt{d(x)}\,\frac{2\dfrac{\partial^2 f_\theta(x)}{\partial\theta_j\partial\theta_i}f_\theta(x) - \dfrac{\partial f_\theta(x)}{\partial\theta_j}\dfrac{\partial f_\theta(x)}{\partial\theta_i}}{\left(f_\theta(x)\right)^{3/2}} \quad , \qquad i, j=1, 2, \ldots, 2k-1$$

and $\mathbf{I(\theta)}$ is the Fisher information matrix. Note that the variance-covariance matrix tends to the inverse of Fisher information matrix as the sample size increases. To see this, it suffices to show that, for large sample size, the matrix $\mathbf{H}$ tends to the Fisher information matrix. This is so since the quantity under the root in equations (5.8) and (5.9) tends to 1 and, hence, the derivatives of these equations are the same as the entries of the Fisher information matrix. So, if $\mathbf{H} \to \mathbf{I}(\theta)$, the matrix $\mathbf{V}$ tends to the inverse of the Fisher information matrix, i.e. to the variance covariance matrix of the ML estimator.

A simulation experiment was carried out in order to assess the normality of the estimates. 1000 samples of sizes n=50,100,500 were drawn from a well separated 2-finite Poisson mixture with parameter vector $\mathbf{\grave{e}}$=(0.5, 1, 10) and a mixture with components closer together with $\mathbf{\grave{e}}$=(0.5, 1, 3). Figures 5.1 and 5.2 depicts the Probability-Probability plots for these cases. A straight line indicates normality of the distribution. It can be seen that for samples of size n >100 the normality is evident. However, for smaller sample sizes, deviations from normality are present. Note also that for well separated components the normality is clearer than for cases with

components close together. This experiment reveals that the estimates tend to be normal depending on both the sample size and the distance between the components. Components close together tend to produce estimates which approach normality slower.

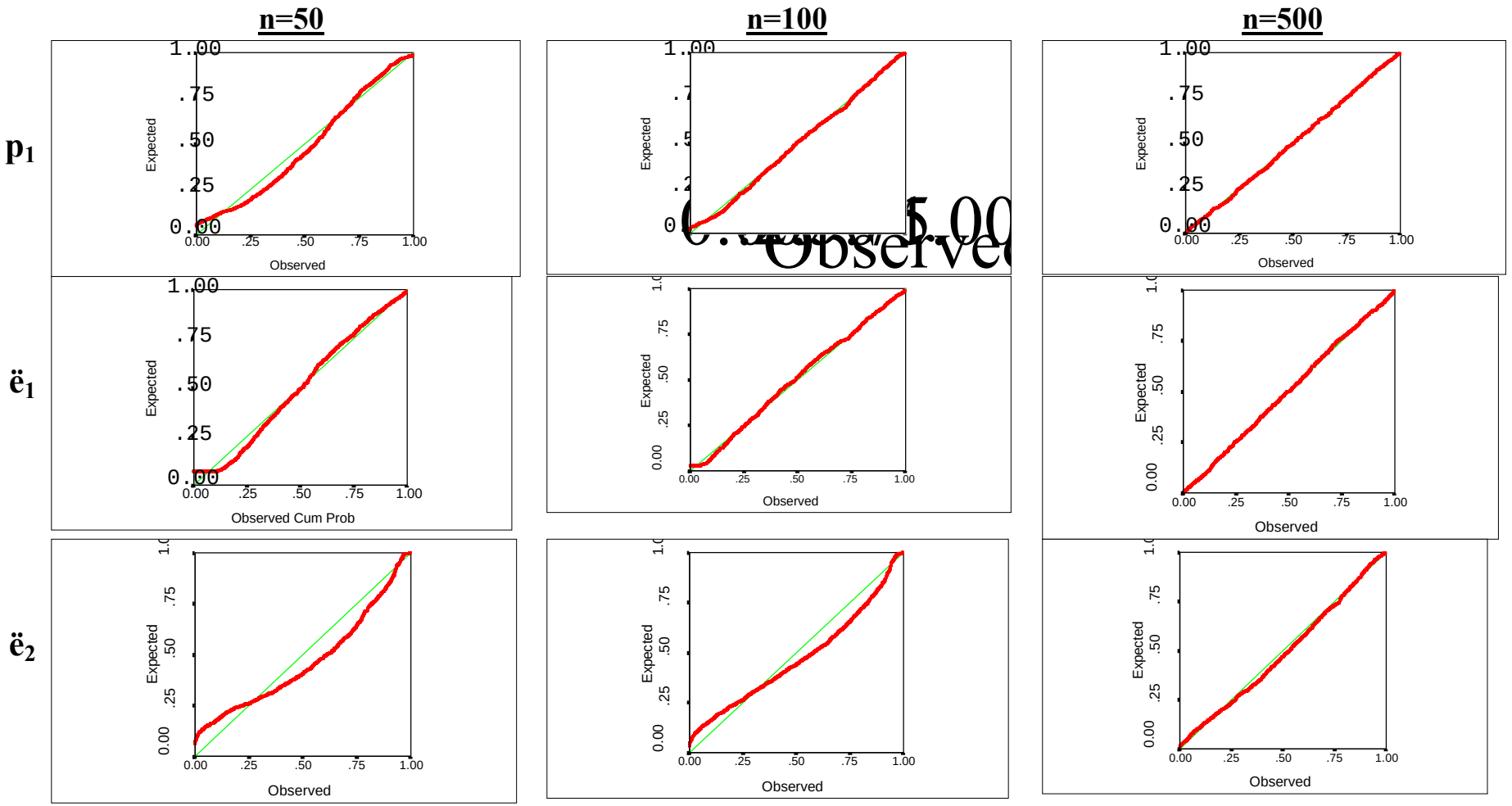


Figure 5.1 Normal P-P plots for p_1 (a-c), $\ddot{e}_1$ (d-f) and $\ddot{e}_2$ (g-i). 1000 samples of size n were drawn from a 2-finite mixture with parameters 0.5,1,3. The sample sizes used were n=50, 100, 500. It is evident that as the sample size increases the estimators tend to normality.

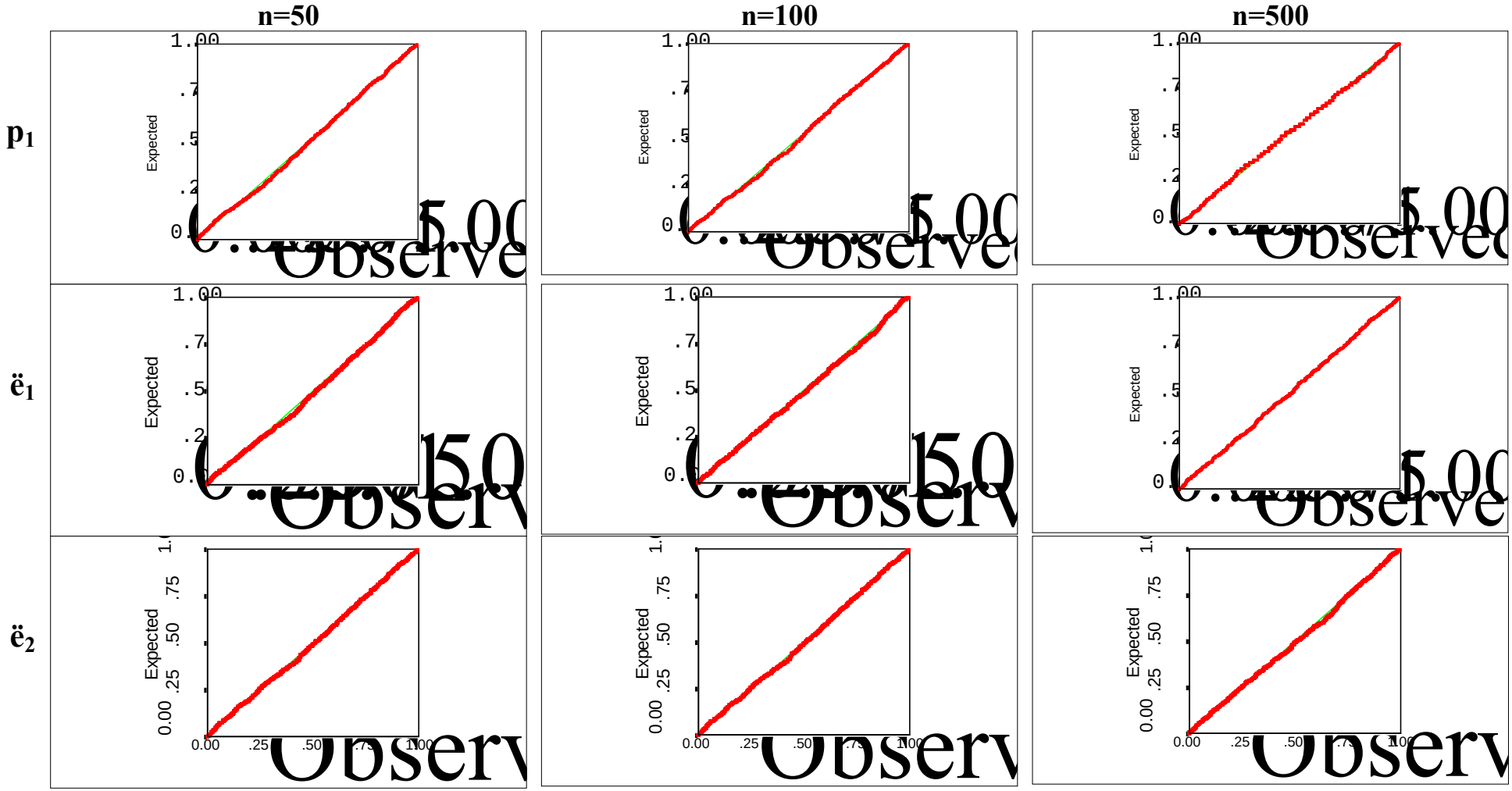


Figure 5.2 Normal P-P plots for p_1 (a-c), $\ddot{e}_1$ (d-f) and $\ddot{e}_2$ (g-i). 1000 samples of size n were drawn from a 2-finite mixture with parameters 0.5,1,10. The sample sizes used were n=50, 100, 500. It is evident that as the sample size increases the estimators tend to normality.

5.5 Measures of robustness and breakdown points

In examining the robustness of some procedures certain measures are needed. A common such measure is the so-called Influence Function (IF) defined by

$$IF(x,T,F)=\lim_{t\downarrow 0}\frac{T\left((1-t)F+t\Delta_x\right)-T(F)}{t} \quad , \tag{5.10}$$

whenever this limit exists (see, e.g., Hampel *et al.*, 1986). Here *T(F)* is a functional based on the distribution function *F*, which is usually the empirical distribution function of the data and Δ_x is a degenerate distribution at x. For example, for the MHD estimator the functional *T(F)* is defined as *T(F)* = $\{\theta\in\mathbf{\Theta}: D\left(d,f_\theta\right)$ is minimised $\}$, where **È** is the parameter space. The importance of the influence function lies in its heuristic interpretation: it describes the effect on the estimate of an infinitesimal contamination at the point x standardised by the mass of contamination.

In many instances the IF is very hard to be calculated and thus some other versions are more appropriate. One of them is to use the empirical counterpart of the IF, namely the Empirical Influence Function (EIF). According to Hampel *et al*. (1986, pp. 93), the EIF of the estimator based on any sample, is a plot of the values of the estimator if one more observation (contaminant) is added at the point x. So, in Figure 5.3 we can see the EIF of the sample in example 5.1 for the 2-finite Poisson mixture and its estimates for both the ML and the MHD methods of estimation . Each time we contaminated our initial sample of size 25 by adding to it a 26th observation at the point x, (x=0,1,2...,20) and calculated both the ML estimates and the MHD estimates of the parameters on the resulting contaminated sample of size 26. Table 5.4 contains the estimates for each case.

Table 5.4
Estimates for the parameters of a 2-finite Poisson mixture based on the dataset of example 5.1 and on the same sample when a new observation is added with value x

	MHD			ML		
	p_1	$\ddot{e}_1$	$\ddot{e}_2$	p_1	$\ddot{e}_1$	$\ddot{e}_2$
initial sample (no contamination)	0.437	0.347	3.003	0.436	0.402	3.308
contamination at x=						
0	0.452	0.293	2.981	0.445	0.335	3.269
1	0.506	0.521	3.112	0.508	0.578	3.469
2	0.401	0.321	2.875	0.395	0.369	3.130
3	0.350	0.216	2.832	0.369	0.288	3.126
4	0.407	0.311	3.013	0.390	0.332	3.259
5	0.430	0.355	3.124	0.417	0.392	3.418
6	0.450	0.388	3.186	0.448	0.459	3.598
7	0.485	0.483	3.588	0.480	0.531	3.800
8	0.502	0.501	3.562	0.509	0.596	4.008
9	0.498	0.482	3.440	0.536	0.652	4.219
10	0.484	0.447	3.296	0.562	0.708	4.447
11	0.470	0.416	3.185	0.584	0.756	4.673
12	0.459	0.393	3.112	0.603	0.799	4.897
13	0.454	0.382	3.073	0.620	0.838	5.121
14	0.450	0.375	3.052	0.636	0.874	5.343
15	0.449	0.371	3.040	0.650	0.909	5.571
16	0.448	0.368	3.034	0.664	0.947	5.808
17	0.448	0.370	3.034	0.678	0.987	6.057
18	0.448	0.370	3.033	0.695	1.038	6.342
19	0.448	0.371	3.034	0.717	1.112	6.711
20	0.448	0.371	3.034	0.961	2.039	19.98

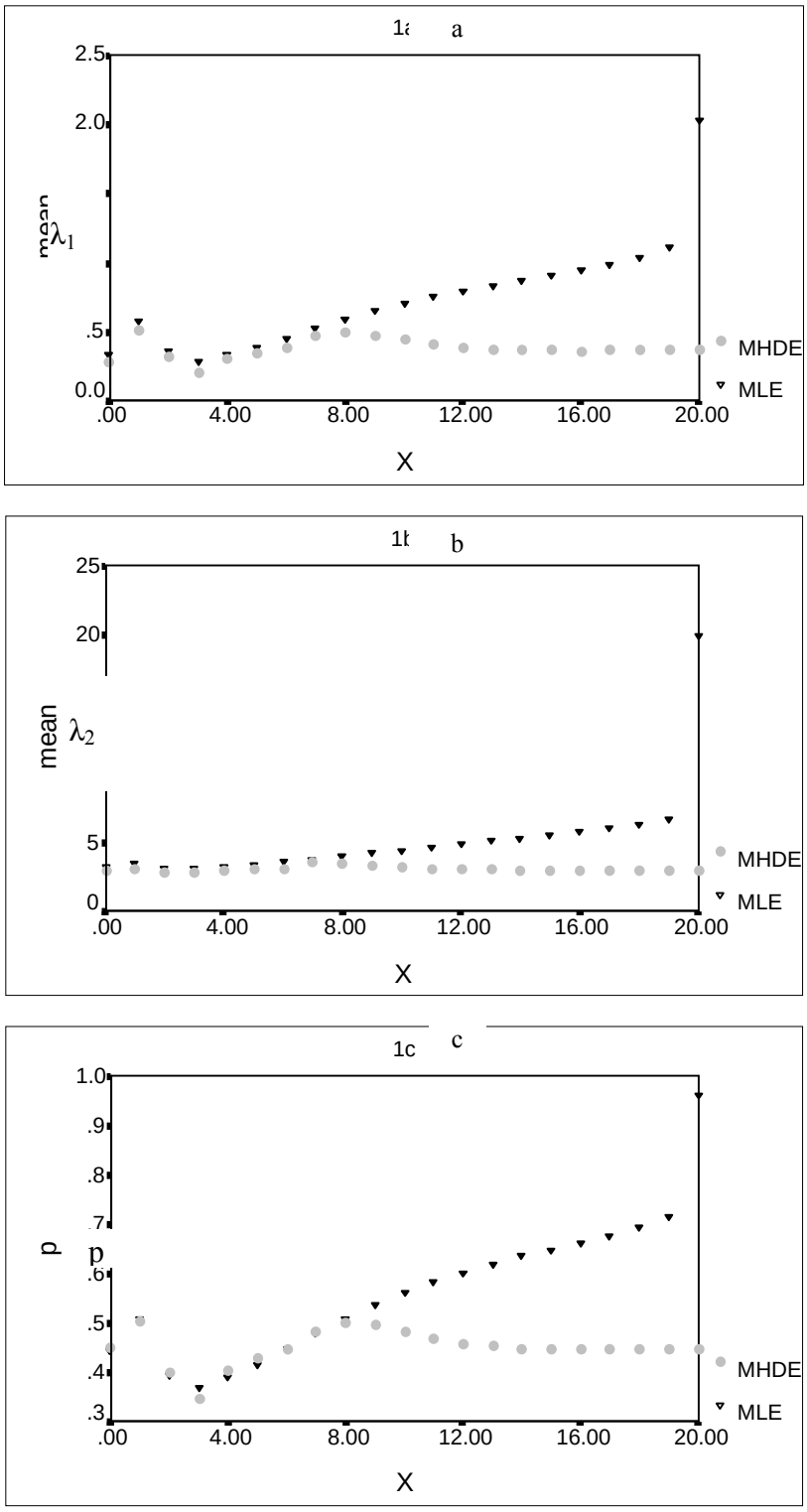


Figure 5.3 The Empirical Influence Function for the parameters of a 2-finite Poisson mixture fitted to the data in Table 5.2. Figures a-c depict the function for $\ddot{e}_1$, $\ddot{e}_2$ and p_1 respectively. The MHD method seems to be more robust to the presence of an outlier in the data.

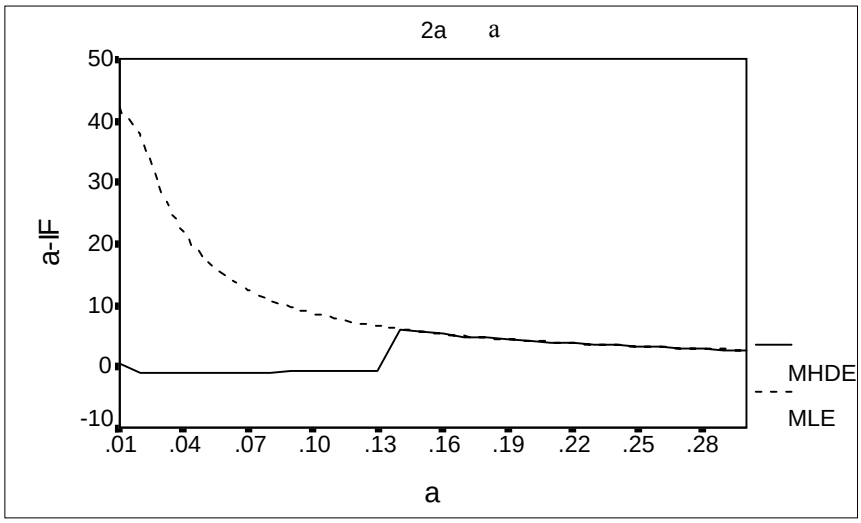
From Figure 5.3 one can see that the MHD estimates are not influenced so much by the addition of one more observation, especially at points far from the main body of the sample. It is interesting how stable the MHD estimator remains for x>10; it seems that it entirely ignores the new observation. In the sequel, we provide an explanation of why this happens. Jorgensen (1990) proposed the use of the EIF as a diagnostic tool for the influence of an observation in finite mixture models. He also reported the influence of observations far from the main body of the data to the ML method. His results are quite similar to those obtained in the present section as far as the influence of an outlier to the ML estimates is concerned.

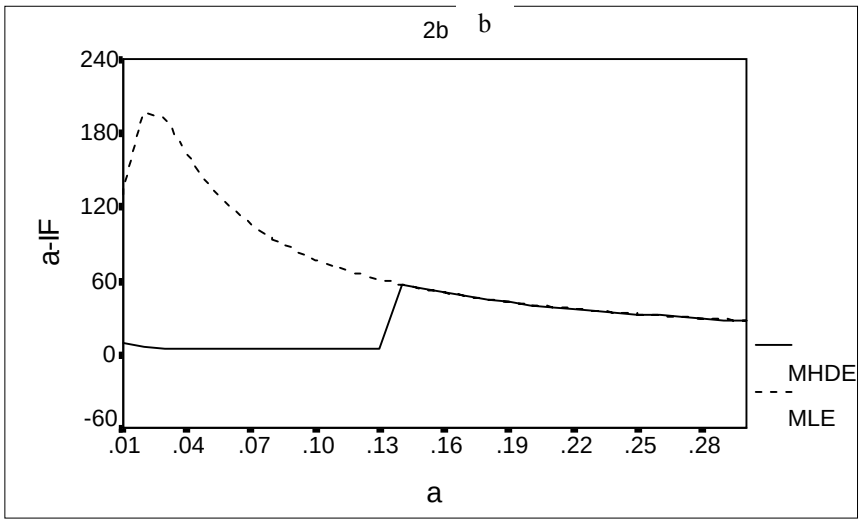
An alternative measure of robustness is the so called á-Influence Function, (Beran, 1977). This measures the change in the estimators if we add one more component in the model and we assign to it a probability equal to á. In particular, the á-Influence Function (á-IF) is defined as:

$$a - IF(x, T, F) = \lim_{t \downarrow 0} \frac{T\left((1-a)F + ag_z\right) - T(F)}{a} \quad , \qquad (5.11)$$

where g_z is one more component of the same distribution (Poisson in our case), with parameter z.

The difference from the simple IF is that the simple IF measures the influence of one more observation at the point x, while the á-IF measures the influence of one more component with mixing proportion á. To illustrate this, consider the model 0.5 *Po*(1) + 0.5 *Po*(3) (i.e., an equiprobable mixture of a Poisson distribution with parameter equal to 1 with a Poisson distribution with parameter equal to 3) and also consider a *Po*(12) distribution (z=12) as the contaminant g_z. Analytic evaluation of the á-IF is not possible. So, we calculate the á-IF numerically for both the MHD method and the ML methods of estimation, when we add a new component (a *Po*(12) distribution) with mixing proportion á. The results are depicted in Figure 5.4. Table 5.5 contains the values of the estimates for various values of the probability á assigned to the new component.





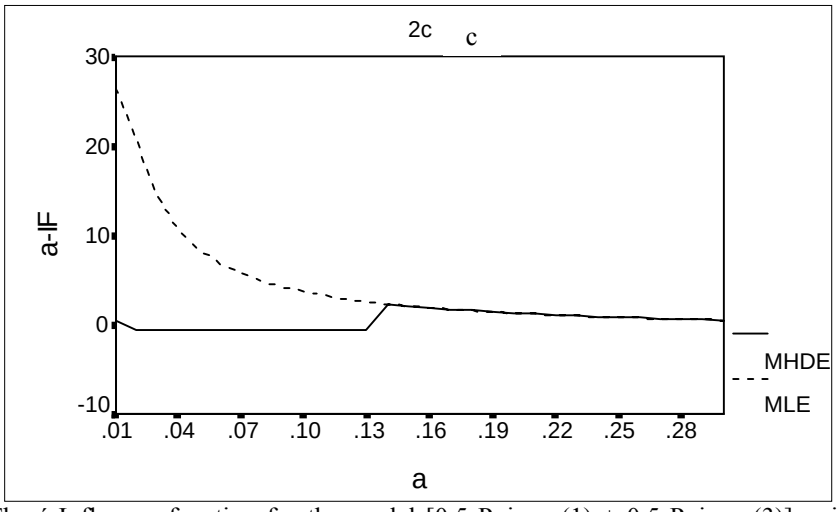


Figure 5.4 The á-Influence function for the model [0.5 Poisson(1) + 0.5 Poisson(3)] with a Poisson (12) distribution as the contaminant. Figures a-c depict the function for $ë_1$, $ë_2$ and p_1 respectively. When á is small, the ML method is influenced very much. Also, the influence is larger for the parameter $ë_2$. The MHD method seems to be more robust to contamination.

Table 5.5
Estimates of the parameters of a 2-finite Poisson mixture with vector of parameters (0.5,1,3) when a new component, a Po(12) distribution, is added with probability α

	MHD estimates			ML estimates		
	p_1	$ë_1$	$ë_2$	p_1	$ë_1$	$ë_2$
no contamination	0.5	1	3	0.5	1	3
addition of Po(12) component with probability α=						
0.01	0.506	1.005	3.096	0.765	1.421	4.300
0.02	0.491	0.980	3.119	0.916	1.758	6.958
0.03	0.485	0.972	3.161	0.935	1.846	8.766
0.04	0.479	0.964	3.202	0.931	1.865	9.528
0.05	0.473	0.956	3.243	0.924	1.872	9.948
0.06	0.467	0.948	3.283	0.915	1.874	10.219
0.07	0.462	0.941	3.325	0.905	1.874	10.413
0.08	0.456	0.934	3.368	0.895	1.873	10.555
0.09	0.451	0.928	3.413	0.885	1.872	10.674
0.10	0.447	0.923	3.461	0.875	1.870	10.766
0.11	0.443	0.918	3.513	0.866	1.869	10.844
0.12	0.439	0.915	3.569	0.856	1.867	10.911
0.13	0.436	0.912	3.630	0.845	1.865	10.967
0.14	0.835	1.854	10.964	0.836	1.864	11.020
0.15	0.825	1.852	11.007	0.825	1.862	11.063
0.16	0.815	1.850	11.045	0.815	1.860	11.105
0.17	0.805	1.849	11.078	0.805	1.858	11.139
0.18	0.795	1.847	11.109	0.795	1.856	11.173
0.19	0.784	1.845	11.137	0.785	1.855	11.203
0.20	0.774	1.843	11.163	0.775	1.853	11.229
0.21	0.764	1.841	11.186	0.765	1.851	11.255
0.22	0.754	1.839	11.208	0.756	1.849	11.278
0.23	0.744	1.838	11.228	0.745	1.847	11.299
0.24	0.734	1.836	11.247	0.736	1.845	11.319
0.25	0.724	1.834	11.265	0.726	1.844	11.339
0.26	0.715	1.832	11.281	0.716	1.842	11.357
0.27	0.705	1.830	11.297	0.706	1.840	11.374
0.28	0.695	1.828	11.312	0.696	1.838	11.388
0.29	0.685	1.827	11.326	0.686	1.837	11.404
0.30	0.675	1.825	11.339	0.676	1.835	11.418

Again, it is clear that the MHD estimates are more conservative in accepting this new component. So, their behaviour is more stable than that of the ML estimates. It is interesting to note that when á increases the difference between the two methods decreases. An intuitive explanation of this is that the MHD method is "persuaded" that the new observation is not an outlier but it comprises one more component. It is also interesting to note how the behaviour of the two methods of estimation changes with á, a fact that will again be seen in the simulation comparison. For small á (low contamination) the MHD estimates are far better, but with á increasing, the two methods work in the same manner. Concluding, for small amounts of contamination the MHD method of estimation seems to be preferable. In the sequel, we examine this aspect in detail. Note that usually one treats as outliers, a few observations far from the main body of the data. In other words, one may have to regard a small fraction á of the observations as outliers. If the proportion of spurious observations is high, then clearly these cannot be regarded as outliers. It is interesting to point out the jump of the á-IF of the MHD method at the point á=0.13. This can be considered as the breakdown point in the sense of Simpson (1987).

5.6 The Algorithm HELMIX for Deriving the Minimum Hellinger Estimates

In many instances, the applicability of an estimation method is mainly the result of the existence of efficient and easily applicable algorithms for calculating the estimates rather than the results of their properties. Hence the MHD method proposed above needs to be accompanied by an algorithm which can be used easily. In the sequel, such an algorithm termed as the HELMIX algorithm which is fairly easy to be programmed in any computer is provided. It is an iterative algorithm, similar in nature to the EM algorithm for ML estimation in finite mixture models.

The algorithm is developed using the estimating equations given in section 5.3. From (5.8) one obtains, using the recurrence relation $f(x,\lambda) = f(x-1,\lambda)\, x/\lambda$ for the Poisson probabilities, that

$$\sum_{x=0}^{\infty} \sqrt{d(x)} w_{xj} \left(x - \lambda_j \right) = 0 \qquad , \qquad j=1,\ 2,\ \ldots,\ k,$$

where $w_{xj} = \dfrac{f(x,\lambda_j)}{\sqrt{f_\theta(x)}}$.

Solving these equations with respect to the parameters ëj, *j = 1, 2, . . . , k* one obtains

$$\lambda_j = \frac{\sum_{x=0}^{\infty} w_{xj} x \sqrt{d(x)}}{\sum_{x=0}^{\infty} w_{xj} \sqrt{d(x)}} \quad , \qquad \text{j=1, 2, . . . , k} \qquad (5.12)$$

i.e. , the MHD estimates are weighted versions of the sample mean. Note that a similar result was obtained for the ML estimates, but with different weights (see equation (3.15) in section 3.2).

From (5.7) we obtain

$$\sum_{x=0}^{\infty} \sqrt{\frac{d(x)}{f_\theta(x)}} f(x,\lambda_i) = \sum_{x=0}^{\infty} \sqrt{\frac{d(x)}{f_\theta(x)}} f(x,\lambda_k) \quad . \qquad (5.13)$$

Also, multiplying the i-th equation in (5.7) by p_j and adding the resulting equations yields

$$\sum_{x=0}^{\infty} \sqrt{\frac{d(x)}{f_\theta(x)}} f(x,\lambda_k) = \sum_{x=0}^{\infty} \sqrt{d(x) f_\theta(x)} \quad . \qquad (5.14)$$

The last two equations (5.13) and (5.14) lead to

$$\sum_{x=0}^{\infty} \sqrt{d(x) f_\theta(x)} = \sum_{x=0}^{\infty} \sqrt{d(x)} w_{xj} \quad , \qquad j=1, 2,. . ., k. \qquad (5.15)$$

Following Behboodian (1970), we may multiply them by p_j obtaining

$$p_j = \frac{\sum_{x=0}^{\infty} p_j w_{xj} \sqrt{d(x)}}{\sum_{x=0}^{\infty} \sqrt{d(x) f_\theta(x)}} \quad , \qquad j=1, 2, . . . , k \ . \qquad (5.16)$$

Equations (5.11) and (5.15) are the basis of our iterative scheme. Note that this derivation is very similar to that introduced by Behboodian (1970) for the derivation of the ML estimates in the case of finite normal mixtures.

Hence the algorithm can be described with the following steps:

Step 1: Given the values obtained from the i-th iteration $\lambda_j^{(i)}$, and $p_j^{(i)}$, *j=1, ..., k,* calculate the weights w_{xj}, using

$$w_{xj} = \frac{f(x|\lambda_j^{(i)})}{\sqrt{f_\theta(x)}} \quad , \tag{5.17}$$

where $f_\theta(x)$ is calculated using the estimates from the i-th iteration.

Step 2: Calculate the new parameter estimates using

step 2a

$$\lambda_j^{(i+1)} = \frac{\sum_{x=0}^{m} w_{xj} x \sqrt{d(x)}}{\sum_{x=0}^{m} w_{xj} \sqrt{d(x)}} \quad , \qquad \text{j=1, 2, . . . ,k} \tag{5.18}$$

step 2b

$$p_j^{(i+1)} = \frac{\sum_{x=0}^{m} p_j^{(i)} w_{xj} \sqrt{d(x)}}{\sum_{x=0}^{m} \sqrt{d(x) f_\theta(x)}} \quad , \qquad \text{j=1, . . . , k} \tag{5.19}$$

where m denotes the largest observed value.

Step 3: Check if some convergence criterion is satisfied, otherwise go back to step 1, using the current estimates as initial values to make the next iteration.

In the sequel, we refer to this algorithm as the HELMIX algorithm.

Clearly, we only need some initial values for the estimates. If the initial values are within the acceptable range for the parameters, the estimated values are also within the acceptable range of parameters.

HELMIX is similar to the well known EM algorithm for ML estimation of finite mixture models described in chapter 3. The only difference is the calculation of the weights in step 1. If we use $w_{xj} = f(x|\lambda_j^{(i)}) / f_\theta(x)$ as weights, i.e. if instead of taking the root of $f_\theta(x)$, we take $f_\theta(x)$ and the observed frequencies themselves, then HELMIX reduces to the EM algorithm for ML estimation for finite mixture models. Table 5.6 contains the details for the two algorithms.

Table 5.6
Description of the EM algorithm for ML estimation and the HELMIX algorithm for MHD estimation

	MHD method	ML method
E-step	$w_{xj} = \dfrac{f(x\|\lambda_j^{(i)})}{\sqrt{f_\theta(x)}}$	$w_{xj} = \dfrac{f(x\|\lambda_j^{(i)})}{f_\theta(x)}$
M-step	$\lambda_j^{(i+1)} = \dfrac{\sum_{x=0}^{m} w_{xj} x\sqrt{d(x)}}{\sum_{x=0}^{m} w_{xj}\sqrt{d(x)}}$	$\lambda_j^{(i+1)} = \dfrac{\sum_{x=0}^{m} w_{xj} x d(x)}{\sum_{x=0}^{m} w_{xj} d(x)}$
	$p_j^{(i+1)} = \dfrac{\sum_{x=0}^{m} p_j^{(i)} w_{xj}\sqrt{d(x)}}{\sum_{x=0}^{m}\sqrt{d(x) f_\theta(x)}}$	$p_j^{(i+1)} = \dfrac{\sum_{x=0}^{m} p_j^{(i)} w_{xj} d(x)}{\sum_{x=0}^{m} d(x)}$

Lindsay (1994) described an iterative algorithm for minimum distance estimation for discrete one parameter exponential families. Our algorithm reduces to the one described by Lindsay (1994) if we set k=1 (simple Poisson case). A similar reweighted algorithm can be found in Basu and Lindsay (1994) for MHD estimation for continuous models.

Our algorithm seems to share some common properties with the EM algorithm for ML estimation in the case of mixture models. These are the slow convergence and the dependence on the choice of the initial values. Again if the true values are known, these are very successful initial choices as in the case of the ML method. For all our simulations, the algorithm converged to a minimum. As in the case of the EM algorithm, the attained minimum might not be a global one. A good strategy is to start from several different initial values so as to ensure that the global minimum is obtained.

It is worth mentioning that the above described derivation of the algorithm might be generalised for a broad family of distances. Thus, depending on the weight function, several other distances might be minimised by applying similar iterative algorithms.

5.7 An Application

To illustrate the appropriateness of the MHD methods of estimation for data sets prone to outliers in the case of k-finite Poisson mixtures, consider the data in Table 5.7. They concern the number of environmental complaints per day placed by phone in an environmental station for the year 1985 in Nederlands. The data were

kindly provided by Prof. Paul Eiler. The high overdispersion of the data makes the use of a mixed Poisson distribution appropriate to model the number of environmental complaints. The mean is 22.11 while the variance is 324.08 (almost 15 times higher than the mean). A simple Poisson model is quite inappropriate because of this overdispersion. Moreover, the data are highly skewed, with a very long right tail. The ML estimators are expected to be influenced by the data at the tail and thus the estimates will not be a reasonable choice if we want to use them for describing the situation. So, the MHD method of estimation may be more appropriate as it seems not to be affected so much by the observations at the right tail.

To these data a 3-finite Poisson mixture was fitted using both the ML method and the MHD method of estimation for comparison purposes. Table 5.7 contains the expected frequencies using both the methods while Table 5.8 contains the parameter estimates. The choice of a model with 3 components was made mainly for illustrative purposes.

Table 5.7
Observed and expected frequencies of environmental complaints placed in an environmental station in 1985

x	observed frequencies	expected frequencies		x	observed frequencies	expected frequencies	
		MHDE	MLE			MHDE	MLE
0-4	37	22.95	4. 70	45-49	11	7.20	2.19
5-9	67	96.71	85.71	50-54	3	1.85	0.30
10-14	69	60.75	114.47	55-59	3	0.30	0.13
15-19	56	70.20	26.19	60-64	7	0.03	0.49
20-24	28	37.76	8.68	65-69	2	0.01	1.47
25-29	23	12.17	25.19	70-79	3	0	7.64
30-34	21	16.08	37.42	80-89	1	0	9.05
35-39	13	22.19	26.90	90-99	2	0	3.65
40-44	13	16.78	10.20	≥100*	6	0	0.60

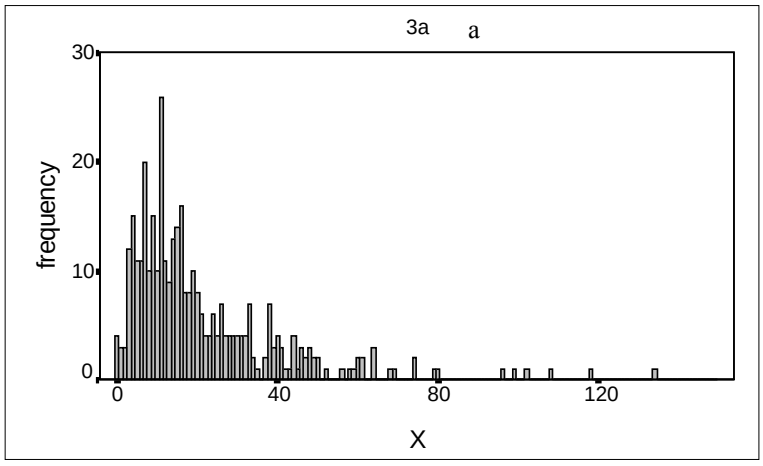
* the actual observations were (102,108,118,134,158,185)

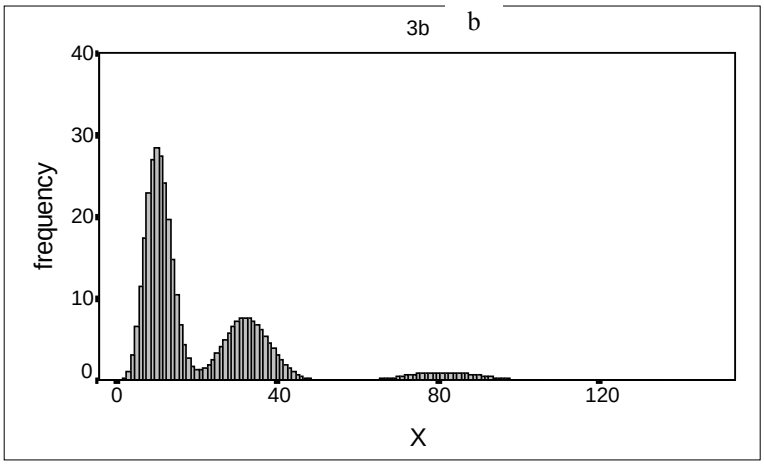
Table 5.8
The parameter estimates for both the methods for the data in Table 5.7

	p_1	p_2	$ë_1$	$ë_2$	$ë_3$
MHDE	0.390	0.418	7.136	17.331	37.676
MLE	0.635	0.302	10.559	32.587	81.423

The method of moments was also applied, but it failed to give us reasonable estimates ($ë_1 < 0$). Figure 5.5 depicts the observed frequencies and the fitted frequencies for both methods. From Table 5.7 it can be observed that the distribution fitted by the ML

method of estimation has a heavier right tail than that of the observed distribution with a bump in the range 70-89. The distribution fitted by the MHD method, on the contrary, has a smoother right tail and it provides a relatively better fit to the data. What is, however, important to note is not the fit itself but the fact that the ML method tries to fit a component at the tail and hence the tail influences the estimation. On the other hand, the MHD method is more conservative in the sense that it treats the right tail quite differently. The ML method is influenced by the very large values at the tail and thus it tries to fit a component to these observations. Quite to the contrary, the MHD method seems to ignore these observations. These high values may be outliers, for example, some days with unexpectedly high number of phonecalls. So, it is reasonable to handle these outliers with care. The great difference in the two estimates demonstrates precisely, how the choice of an estimation method can affect the results. Of course, assessing whether the right tail of a data set can be attributed to the presence of some kind of contamination would enhance the practical value of the MHD method. An interesting interpretation of why this phenomenon of markedly different estimates occurs is given in the next section.





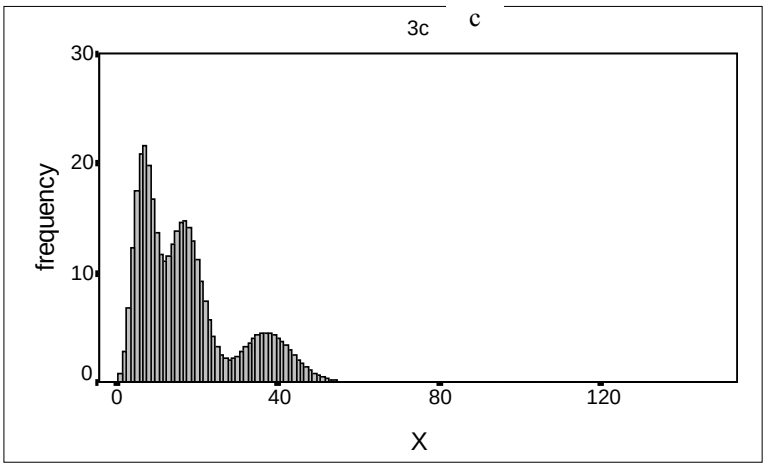


Figure 5.5 Histograms of the observed frequencies (a), the expected frequencies via the ML method (b) and the expected frequencies via the MHD method (c) for the data of Table 5.7.

Concluding, we may say that for data with a long right tail that can be attributed to ″unexpected″ or even ″unreasonable″ values (due to some mistakes in the collection of data) the ML method must be used with caution while the MHD method offers itself as a very interesting alternative.

5.8 Comparison of the MHD Method to the ML Method

5.8.1 General Comments

In parametric estimation two fundamental - but potentially competing- aspects become of interest: The aspect of efficiency when the model has been appropriately specified and the aspect of robustness when it has not. Unfortunately, satisfying both is very difficult and a trade-off between them is thus necessary. Ôhe MHD method for finite Poisson mixtures and the ML method are compared in the sequel with respect to both aspects.

In general, for the parameter θ_i , the estimating equation in the case of the ML method is given by

$$\sum_{x=0}^{\infty} \frac{d(x)}{f_\theta(x)} \frac{\partial f_\theta(x)}{\partial \theta_i} = 0 \quad , \tag{5.20}$$

while, in the case of the MHD method, it is given by

$$\sum_{x=0}^{\infty} \sqrt{\frac{d(x)}{f_\theta(x)}} \frac{\partial f_\theta(x)}{\partial \theta_i} = 0 \quad . \tag{5.20}$$

Equations (5.19) and (5.20) are useful for comparison purposes. Clearly, if the model is well specified and the sample size is large $(n \to \infty)$ the quantity under the square root must be close to 1 and, hence, the square root of this quantity is itself close to 1; it is expected that the two methods will behave similarly. In the ideal case of exact specification of the model, the ratio $\frac{d(x)}{f_\theta(x)}$ equals 1 for every x and the two methods coincide. On the other hand, for values of x for which the ratio $\frac{d(x)}{f_\theta(x)}$ is high (as in the case of outliers) the MHD method gives less weight to the observation, the estimation being thus not so sensitive to outliers. As a result, the MHD method works better with datasets prone to outliers. Simpson (1987) showed that for large values of x, an improbable count has a smaller impact on the MHD estimates than on the ML estimates. Our results, referring to example 5.1, on how a new observation far away from the bulk of data influences the two methods are not at variance with Simpson's findings.

It is worth mentioning the similarity of the estimating equations given in (5.20) to the so-called weighted maximum likelihood equations (see, e.g., Field and Smith, 1994, Markatou, 1996, Markatou *et al.*, 1997, Markatou, 1998). In these papers, weighted likelihood methods are treated so as to provide robust estimates. The MHD estimating equations can be considered as weighted likelihood equations.

Another intuitive interpretation for the behaviour of the two methods might be traced in the following. Suppose that we have an outlier observation x. This observation contributes via the logarithm of its probability function to the loglikelihood function, while in the case of the MHD method it contributes via the square root of its probability function. To see this, note that for the ML method the function

$$L(\theta) = \sum_{x=0}^{\infty} d(x) \ln\left(f_\theta(x)\right) \tag{5.22}$$

must be maximized, while for the MHD method it can be seen from (5.7b) that the function

$$\varphi(\theta) = \sum_{x=0}^{\infty} \sqrt{d(x) f_\theta(x)} \tag{5.23}.$$

must be maximized.

Then, an outlier observation will contribute via its logarithm to the likelihood in the case of the ML method and via its square root in the case of the MHD method since an outlier will have an observed frequency equal to 1. Figure 5.6 depicts the functions $h_1(x) = \ln x$ and $h_2(x) = \sqrt{x}$ as functions of x in the interval (0,1). For very small values of x, close to 0, the behaviour of the logarithm is very rough, tending to very small values quite fast. The derivatives of h_1 and h_2 are $h_1'(x) = \frac{1}{x}$ and $h_2'(x) = \frac{1}{2\sqrt{x}}$, which shows that the function h_1 decreases more rapidly near 0. So, in the case of ML estimation observations with very low probability (small values of x in Figure 5.6) have a very small contribution to the likelihood (small values of ln(x) in Figure 5.6). Since the aim is to maximise the sum of these contributions for all the observations, the method tries to exploit the contribution of observations with low probability, so as to increase the loglikelihood. To do so, observations with low probability are ‘‘made’’ more important by increasing their mass (of course trying to balance with the remaining observations). This can explain why counts with very low probability under the assumed model (the outliers) can influence so much the ML estimates. The square root, on the other hand, is rather flat and this phenomenon does not occur, a fact that offers an explanation for the robustness of the MHD method with respect to outliers. Note also that for the MHD estimation method the contribution of each value x is not proportional to its frequency but it increases more rapidly than in the case of the ML method, as the frequency increases.

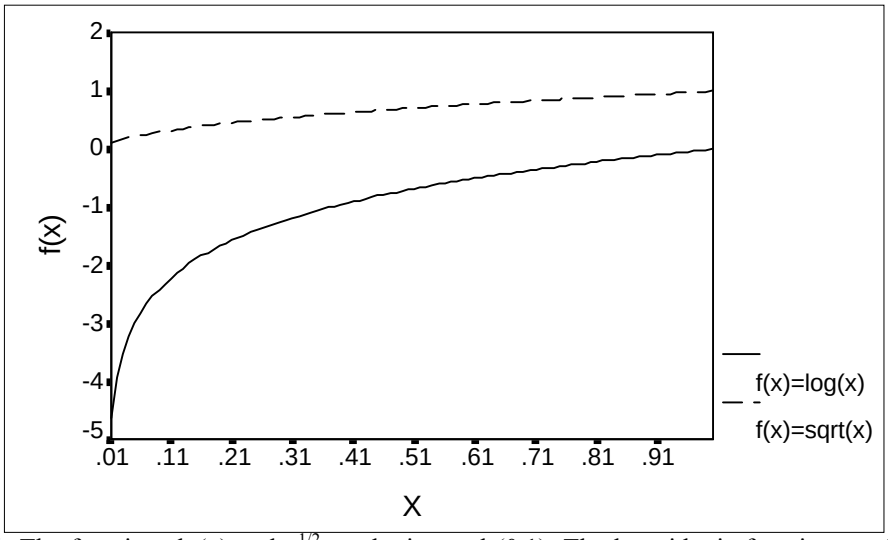


Figure 5.6 The functions ln(x) and $x^{1/2}$ on the interval (0,1). The logarithmic function can be seen to decrease more rapidly near 0.

It is interesting to note that empty cells do not contribute at all in both of the methods. Lindsay (1994) uses the term *inliers* to describe cells with $\frac{d(x)}{f_\theta(x)}$ near 0. These may be regarded as corresponding to cases where observations, which were probable under the model, were not observed. (Note that outliers at the right tail have a large value for $\frac{d(x)}{f_\theta(x)}$). Empty cells are inliers and both methods fail to make use of them. This provides an alternative explanation to that given by Lindsay (1994) on why the two methods cannot treat the inliers. Some authors, (e.g. , Harris and Basu, 1994) have reported that empty cells affect the estimation by the MHD method. Their argument is based on the fact that, by the definition of the Hellinger distance given in (5.7), empty cells (*d(x)*=0) have a non-zero contribution to the distance. Expansion however, as in the second representation of the distance given in (5.23), reveals that this is not true.

The above arguments pertain to a general comparative evaluation of the two methods. For a further discussion the reader is referred to Lindsay (1994).

Let us now focus our attention on the case of mixtures. In robust analysis we consider a contaminated model of the form (1-e) M_1 + e M_2 , where M_1 is the underlying model, M_2 is a contaminant which causes the departure from model M_1 and e is the probability that an observation belongs to the contaminant. Note that the quantity e is itself of practical interest since, as Simpson (1989) and Lindsay (1994) pointed out, for every model there is a value of e which gives an upper bound of possible contamination. Above this point there is a breakdown point for the model. Tamura and Boss (1986) defined this as the value of e which indicates the fraction of the data that can be badly damaged or arbitrarily changed without destroying the estimator.

Clearly, the above model is a mixture and it is true that contaminated models are described as mixture models (e.g. Titterington *et al*., 1985). The contaminated model can be considered as a mixture model with an additional component. So, for example, a contaminated model for a 2-finite Poisson mixture can be considered as a 3-finite Poisson mixture. The question is what the behaviour of the ML method is in

such a situation. The answer is that the ML method usually models the contamination with an additional component. In our case the ML method will work well for a 3-component model, but not for a model with 2-components. Aitkin and Wilson (1980) used the EM algorithm for normal mixture models to model the contamination in a simple normal model. Harris and Basu (1994) did the same for discrete models, comparing the estimates by the MHD method for the Poisson parameter to those by the ML method from a 2-finite Poisson mixture model.

In general, the MHD method is preferable in such cases since it works better if the model is not exact. Simplifying we can say that the MHD methods require more evidence (more observations) to be ″persuaded″ that a new component exists.

The remaining of this section is devoted to simulation comparisons of the two methods. Several sampling schemes will be examined for both robustness and efficiency, starting from the simplest case where only the mixing proportion has to be estimated and treating the more complicated case where all the parameters have to be estimated.

5.8.2 The Case Where Only the Mixing Proportion Must be Estimated

In many situations the parameters of the components are known and then only the mixing proportions must be estimated. This is a simplified estimation problem but it is a good starting point for comparing the two methods.

We will start with such models, assuming that the two components are known and hence only the mixing proportion must be estimated. In this case, the minimisation of the Hellinger distance is easily carried out using the HELMIX algorithm. In such a case we do not update the component parameters and, hence, we iterate between (5.17) and (5.19) keeping the remaining parameters fixed. The same is possible in using the EM algorithm for ML estimation in finite mixture models.

In order to investigate the performance of the MHD method relative to that of the ML method a simulation experiment was carried out. Using several sampling schemes we calculated the MHD estimator and the ML estimator for the mixing proportion. Some of the models were specified as 2-finite Poisson mixtures while some others were contaminated versions of them. The comparison was made in terms of the relative efficiency of the estimates. Alternatively, the ratio of the standard errors

of the estimates (relative standard error) or the ratio of their mean squared errors (relative mean squared error) was used as a measure of relative efficiency.

At first, the efficiency of correctly specified models is examined. Several 2-finite mixtures were considered, for several sample sizes and combinations of population means. Table 5.9 tabulates the relative efficiency (reff) as calculated by the formula:

$$reff = \frac{Var(\hat{p}_{ML})}{Var(\hat{p}_{MHD})} \quad . \tag{5.24}$$

Here $\hat{p}_{ML}$ and $\hat{p}_{MHD}$ are the ML estimator and the MHD estimator for the mixing proportion respectively and $Var(\hat{p}) = \sum_{i=1}^{N} \left(\hat{p}_i - \bar{\hat{p}}\right)^2 \Big/ N$, where $\hat{p}_i$ is the estimated mixing proportion of the i-th sample, $\bar{\hat{p}}$ is the mean of the estimated proportion over all samples and N is the number of replications. Similarly, the mean squared error (MSE hereafter) of an estimate is given by $MSE(\hat{p}) = \sum_{i=1}^{N} \left(\hat{p}_i - p\right)^2 \Big/ N$ where p is the true mixing proportion. All the configurations were replicated 1000 times (N=1000). The sample sizes used were n= 25, 50, 100, 250 and 500 while the mixing proportions were 0.1, 0.25, 0.5, 0.75, 0.9. Table 5.9 summarises the results for the relative efficiency for 2-finite Poisson mixture models. Entries less than 1 favour the ML method while entries with value greater than 1, favour the MHD method.

Table 5.9
Simulation results for estimating the mixing proportion in 2-finite Poisson mixtures when the components are known. The entries are the relative efficiencies based on 1000 replications as calculated by formula (5.24)

p_1	n									
	25	**50**	**100**	**250**	**500**	**25**	**50**	**100**	**250**	**500**
	$ë_1$ =1 $ë_2$=2					**$ë_1$ =1 $ë_2$=3**				
0.10	0.84	0.88	0.91	0.98	0.99	0.80	0.90	0.95	0.98	0.99
0.25	0.75	0.88	0.94	0.98	0.98	0.84	0.91	0.96	0.98	0.99
0.50	0.79	0.87	0.93	0.98	0.99	0.87	0.94	0.94	0.98	0.99
0.75	1.04	0.90	0.92	0.97	0.98	1.07	0.94	0.94	0.96	0.99
0.90	1.62	1.30	0.98	0.93	0.97	1.41	1.19	1.04	0.96	0.96
	$ë_1$ =1 $ë_2$=5					**$ë_1$ =1 $ë_2$=8**				
0.10	0.81	0.88	0.94	0.98	0.99	0.79	0.87	0.91	0.96	0.98
0.25	0.79	0.89	0.95	0.98	0.99	0.84	0.88	0.94	0.97	0.99
0.50	0.84	0.91	0.95	0.97	1.00	0.94	0.93	0.93	0.96	0.99
0.75	0.99	0.91	0.93	0.97	0.99	1.09	1.05	0.98	0.96	0.98
0.90	1.51	1.15	0.99	0.95	0.97	1.30	1.22	1.12	1.02	0.97
	$ë_1$ =2 $ë_2$=3					**$ë_1$ =2 $ë_2$=4**				
0.10	0.75	0.82	0.87	0.94	0.97	0.77	0.85	0.94	0.98	0.98
0.25	0.82	0.86	0.91	0.96	0.98	0.79	0.85	0.93	0.97	0.98
0.50	0.91	0.93	0.92	0.96	0.97	0.84	0.88	0.92	0.97	0.98
0.75	1.05	1.07	1.01	0.95	0.96	0.99	0.95	0.94	0.92	0.97
0.90	1.24	1.33	1.18	1.07	1.04	1.41	1.24	1.15	0.97	0.96
	$ë_1$ =2 $ë_2$=5					**$ë_1$ =2 $ë_2$=8**				
0.10	0.83	0.89	0.92	0.96	0.98	0.87	0.89	0.89	0.96	0.98
0.25	0.77	0.85	0.94	0.97	0.98	0.73	0.85	0.91	0.97	0.98
0.50	0.79	0.87	0.93	0.97	0.99	0.76	0.86	0.93	0.97	0.98
0.75	0.99	0.92	0.93	0.96	0.98	0.97	0.91	0.94	0.97	0.99
0.90	1.44	1.24	1.09	0.94	0.95	1.69	1.23	0.99	0.92	0.96

Looking closely at the results of Table 5.9 one can see that, as was expected, the MHD method tends to the ML method when the sample size increases. On the other hand, for small sample sizes and a large mixing proportion, the MHD estimate is far better than the ML estimate. An explanation for this is the resistance of the MHD method to the presence of spurious observations. It might seem somewhat odd for this to be true for mixing proportions 0.75 and 0.9 only. An explanation might be that that the MHD method ignores the outliers in favour of the more probable observations. So, when the mixing proportion is small, the second component dominates the mixture. Keeping in mind that the estimated mixing proportion was the one with the smallest mean , it is obvious that outliers can occur only at the right tail and hence near the other component. So, a large value of the mixing proportion implies that the majority

of the observations comes from the first component, which is closer to 0, and only a few observations come from the second component which may generate observations at the right tail of our dataset.

So, from the results on the case of 2-finite Poisson mixture models (exact models) the MHD method seems to be efficient for large samples as well as for small samples with a large mixing proportion.

Let us now look at contaminated models. We considered three models to investigate the performance of the MHD estimate for the mixing proportion. The models were:

MODEL (1) $(1-\alpha)\left[pPo(1)+(1-p)Po(3)\right]+\alpha Po(7)$,

where p is the mixing proportion to be estimated and á is the amount of contamination, namely the probability assigned to the third added component (the contaminant) which was a Poisson variable with parameter equal to 7. The values used for á were 0.005, 0.01 and 0.05 . The values of p used in the simulations were 0.1, 0.25, 0.5, 0.75 and 0.9.

MODEL (2) $(1-\alpha)\left[pPo(1)+(1-p)Po(3)\right]+\alpha Po(12)$.

This is similar to MODEL (1), only now the new component is a Po(12) variable.

MODEL (3) $(1-\alpha)\left[pPo(1)+(1-p)Po(5)\right]+\alpha Po(15)$.

Here the contamination amount á was assumed to take values 0.01, 0.05,0.1. The relative efficiencies and the relative MSEs are reported in Table 5.10, based on 1000 replications.

From Table 5.10 it becomes evident that the MHD estimator behaves better for incorrectly specified models. Even for a very low contamination it appears to be more robust. Further, for large values of p the MHD estimator appears again to be more efficient. Another interesting feature is the rapid increase of the relative MSE. An explanation for this might be the tendency that the MHD estimator exhibits to underestimate the mixing proportion for small sample sizes. As the sample size increases the bias is reduced and hence the MSE is reduced leading to a great improvement in accuracy as compared to the ML estimator. Again, we may interpret the difference in the behaviour of the MHD and ML estimators for small mixing

proportions as the result of the fact that count data are constrained on the positive axis so that outliers may occur only at the right tail.

Table 5.10
Relative efficiencies (RE) and relative MSEs (RM) for estimating the mixing proportion based on 1000 replications

		α=0.005					α=0.01					α=0.05				
	n	**25**	**50**	**100**	**250**	**500**	**25**	**50**	**100**	**250**	**500**	**25**	**50**	**100**	**250**	**500**
p_1		**MODEL 1**														
0.1	**RE**	0.80	0.92	0.95	0.98	0.99	0.84	0.90	0.95	0.98	0.99	0.84	0.88	0.94	0.97	0.97
	RM	0.58	0.81	0.89	0.97	0.99	0.64	0.80	0.90	0.98	1.00	0.64	0.79	0.94	1.09	1.14
0.25	**RE**	0.86	0.92	0.95	0.98	0.99	0.83	0.92	0.96	0.98	0.99	0.82	0.91	0.95	0.97	0.98
	RM	0.67	0.79	0.89	0.95	0.98	0.61	0.85	0.93	0.98	1.02	0.67	0.93	1.07	1.19	1.26
0.5	**RE**	0.88	0.91	0.96	0.98	0.99	0.88	0.92	0.95	0.98	0.99	0.83	0.89	0.94	0.98	0.99
	RM	0.71	0.72	0.90	0.96	0.99	0.62	0.82	0.91	1.01	1.06	0.78	1.05	1.33	1.41	1.44
0.75	**RE**	1.04	0.94	0.94	0.97	0.96	1.02	0.94	0.92	0.96	0.98	1.00	0.92	0.92	0.96	0.97
	RM	0.80	0.76	0.82	0.94	0.98	0.77	0.78	0.83	1.04	1.20	1.18	1.31	1.63	1.74	1.71
0.9	**RE**	1.40	1.22	1.06	0.94	0.95	1.47	1.19	1.01	0.95	0.98	1.27	1.04	0.95	0.98	0.99
	RM	1.36	1.13	0.89	1.02	1.12	1.45	1.22	1.04	1.34	1.46	2.64	2.58	2.44	2.18	1.99
		MODEL 2														
0.1	**RE**	0.82	0.89	0.95	0.97	0.98	0.79	0.89	0.92	0.97	0.98	0.71	0.83	0.85	0.91	0.93
	RM	0.60	0.75	0.87	0.93	0.99	0.58	0.75	0.85	0.95	1.02	0.46	0.67	0.91	1.13	1.41
0.25	**RE**	0.84	0.92	0.95	0.97	0.98	0.83	0.89	0.94	0.97	0.98	0.82	0.85	0.91	0.93	0.94
	RM	0.64	0.84	0.89	0.94	0.98	0.64	0.78	0.89	0.95	1.02	0.67	0.79	1.03	1.44	1.94
0.5	**RE**	0.88	0.89	0.95	0.96	0.99	0.87	0.89	0.93	0.98	0.98	0.81	0.88	0.92	0.96	0.98
	RM	0.65	0.74	0.84	0.91	1.02	0.69	0.69	0.84	1.05	1.14	0.67	0.99	1.38	2.15	3.18
0.75	**RE**	1.04	0.95	0.92	0.98	0.99	1.02	0.98	0.93	0.96	0.98	1.02	0.93	0.89	0.94	0.96
	RM	0.88	0.74	0.75	0.97	1.07	0.78	0.71	0.81	1.14	1.30	1.17	1.41	2.01	4.06	5.92
0.9	**RE**	1.52	1.28	1.09	0.98	0.96	1.44	1.24	1.05	0.97	0.96	1.60	1.28	1.06	1.05	1.11
	RM	1.32	1.12	0.87	0.96	1.11	1.62	1.26	1.03	1.49	1.96	3.62	3.68	4.91	8.94	11.12
		MODEL 3														
0.1	**RE**	0.80	0.89	0.93	0.97	0.99	0.76	0.82	0.92	0.94	0.95	0.72	0.78	0.88	0.91	0.91
	RM	0.64	0.79	0.87	0.95	0.97	0.59	0.70	0.91	0.99	1.09	0.53	0.65	0.93	1.28	1.52
0.25	**RE**	0.78	0.91	0.95	0.98	0.99	0.72	0.88	0.93	0.94	0.97	0.74	0.82	0.87	0.94	0.93
	RM	0.60	0.80	0.86	0.95	1.00	0.52	0.81	0.95	1.12	1.45	0.61	0.86	1.15	1.80	2.59
0.5	**RE**	0.83	0.89	0.94	0.98	0.99	0.79	0.88	0.92	0.96	0.98	0.77	0.84	0.90	0.94	0.94
	RM	0.55	0.68	0.85	0.92	0.99	0.61	0.77	1.09	1.64	2.31	0.62	1.18	2.01	3.29	4.44
0.75	**RE**	1.03	0.92	0.92	0.95	0.97	0.99	0.93	0.95	1.00	0.98	0.95	0.91	0.92	1.01	0.99
	RM	0.56	0.60	0.70	0.88	0.98	0.95	1.04	1.75	3.19	4.66	1.33	1.80	3.53	7.06	7.52
0.9	**RE**	1.54	1.15	0.99	0.94	0.98	1.72	1.24	1.01	1.03	1.04	1.50	1.25	1.08	1.04	1.09
	RM	0.94	0.75	0.65	0.81	1.07	2.36	2.11	2.90	6.69	9.66	4.23	5.76	8.85	12.65	15.37

5.8.3 The Case Where all the Parameters Have to be Estimated

In this section simulation results for the case where all the parameters have to be estimated are reported. This is the most interesting case for mixture models. The HELMIX algorithm was used for deriving the MHD estimates and the EM algorithm for the ML estimates. We used the same stopping rule for both algorithms. Judging from the results of section (3.3) we stopped iterating when the maximum difference between the parameters of two successive iterations was smaller than 0.0001. Two sets of initial values were used to increase the chance that the obtained maximum (minimum) was not a local extreme. For both methods, the true parameter values and values m ± 0.5, around the sample mean m with equal probability, were considered as initial values. Recall that such initial values were judged as good choices in section (3.3). Again two issues were of interest. The efficiency and the robustness of both methods. We will examine the efficiency using correctly specified models and the robustness using contaminated models.

5.8.3.1 Correctly Specified Models

Let us now examine the case where the model is correctly hypothesised to be a 2-finite Poisson mixture. For each combination of parameters, 4 sample sizes were examined, namely n=50, 100, 250, 500. Each combination was replicated 1000 times.

The usual method for comparing two methods in multiparameter model estimation utilises the ratio of the generalised variances of the estimators. Recall that the generalised variance is the determinant of the variance-covariance matrix. For our case, the generalised variances were computed on the basis of the variance covariance matrices calculated from the simulation. Specifically, from the 1000 replications we calculated the variances for every estimator and the covariances, constructing the variance-covariance matrix. Table 5.11, summarises the results for several 2-finite Poisson mixture distributions. The entries are the values $|V_{ML}|/|V_{MHD}|$, where $|V|$ denotes the determinant of the variance-covariance matrix, and the subscripts indicate the method used. Entries larger than 1 favour the MHD method.

Table 5.11
The ratio of generalised variances of the ML estimator divided by that of the MHD estimator (1000 replications)

n	**50**	**100**	**250**	**500**	**50**	**100**	**250**	**500**
p_1		$ë_1$ **=1**	$ë_2$**=2**			$ë_1$**=1**	$ë_2$**=3**	
0.2	9.293	16.035	6.104	1.952	3.897	5.056	3.390	1.301
0.5	4.725	4.115	3.261	0.929	1.625	1.182	1.164	0.955
0.8	4.238	2.772	1.904	1.113	1.359	1.103	0.806	0.680
		$ë_1$**=1**	$ë_2$**=5**			$ë_1$**=1**	$ë_2$**=8**	
0.2	1.612	1.271	1.098	1.025	0.702	0.890	0.962	0.994
0.5	0.908	0.964	0.976	0.973	0.628	0.821	0.965	0.980
0.8	0.505	0.662	0.866	0.906	0.570	0.645	0.873	0.934
		$ë_1$**=2**	$ë_2$**=4**			$ë_1$**=2**	$ë_2$**=5**	
0.2	7.119	6.317	4.102	1.794	2.579	1.679	1.419	1.103
0.5	2.510	3.062	1.740	0.869	1.140	1.091	0.963	0.937
0.8	2.707	1.865	0.854	0.714	0.952	0.722	0.687	0.660
		$ë_1$**=2.8**	$ë_2$**=3.2**					
0.2	7.607	28.969	22.886	2.213				
0.5	4.771	14.784	5.547	0.954				
0.8	9.286	9.510	4.449	0.844				

A close inspection of Table 5.11 reveals that the ML method works far better for models with well separated components. On the contrary, when the components are close together the MHD method is superior. Hasselblad (1969) has shown that ML estimators have large standard errors when the components are close together. This explains the superiority of the MHD method. In these cases the covariance terms are very large for the ML method resulting in the great superiority of the MHD method. This can be seen from the entries of Table 5.12, where the efficiency for each parameter is not as large as the efficiency based on the generalised variance. Generally, the MHD method performs better for small sample sizes and low mixing proportions. The latter was also shown to be true in the case where only the mixing proportions had to be estimated.

Usually, the ML method works better for well specified models as compared to the MHD method (Lindsay 1994). However, our results constitute a case in which the MHD method performs better. To further examine this remarkable result we report tables with the estimated relative efficiencies for the parameters. The entries are values of the relative efficiency of the MHD method as defined by the ratio of the standard error of the ML estimator divided by the standard error of the MHD estimator.

Table 5.12
Relative efficiencies of the MHD estimators of the parameters of a 2-finite Poisson mixture

	p_1=0.2			p_1=0.5			p_1=0.8		
	p_1	$ë_1$	$ë_2$	p_1	$ë_1$	$ë_2$	p_1	$ë_1$	$ë_2$
n				$ë_1$ =1		$ë_2$=2			
50	1.78	1.00	2.33	1.29	1.00	2.07	1.06	1.05	1.98
100	1.60	0.95	2.72	1.19	0.93	1.93	0.96	0.95	1.87
250	1.32	0.86	2.03	1.00	0.83	1.76	0.82	0.79	1.89
500	1.05	0.76	1.51	0.87	0.75	1.29	0.76	0.73	1.68
				$ë_1$ =1		$ë_2$=3			
50	1.38	0.96	1.81	1.12	0.95	1.53	0.93	0.93	1.41
100	1.44	0.96	2.09	1.03	0.95	1.30	0.86	0.87	1.41
250	1.23	1.00	1.58	0.99	0.97	1.16	0.81	0.82	1.19
500	1.10	0.96	1.13	0.95	0.95	1.03	0.78	0.84	1.06
				$ë_1$ =1		$ë_2$=5			
50	1.27	1.07	1.36	0.96	1.04	1.03	0.72	0.88	0.99
100	1.14	1.10	1.07	0.97	1.02	1.01	0.82	0.95	0.97
250	1.04	1.03	1.02	0.98	1.01	0.99	0.93	0.99	1.00
500	1.01	1.00	1.00	0.99	1.00	0.99	0.96	0.99	0.99
				$ë_1$ =1		$ë_2$=8			
50	0.92	1.00	0.94	0.88	1.01	0.90	0.87	0.99	0.83
100	0.96	1.02	0.97	0.95	1.00	0.95	0.90	0.99	0.89
250	0.99	1.01	0.98	0.98	1.01	0.98	0.97	1.00	0.96
500	0.99	1.01	0.99	0.99	1.00	0.99	0.98	1.00	0.98
				$ë_1$ =2		$ë_2$=4			
50	1.67	0.95	2.17	1.23	0.94	1.89	1.05	1.01	1.77
100	1.58	0.91	2.34	1.12	0.90	1.72	0.94	0.89	1.70
250	1.29	0.87	1.69	1.02	0.91	1.51	0.84	0.79	1.51
500	1.07	0.86	1.32	0.93	0.85	1.13	0.73	0.71	1.17
				$ë_1$ =2		$ë_2$=5			
50	1.50	0.87	1.84	1.06	0.92	1.42	0.92	0.92	1.29
100	1.36	0.93	1.51	1.00	0.94	1.27	0.84	0.84	1.25
250	1.21	1.01	1.19	0.98	0.97	1.08	0.80	0.85	1.10
500	1.07	0.96	1.07	0.96	0.96	1.02	0.80	0.89	1.01
				$ë_1$ =2.8		$ë_2$=3.2			
50	2.20	1.14	2.64	1.62	1.18	2.63	1.27	1.21	2.21
100	2.24	1.05	3.29	1.42	1.11	2.88	1.21	1.21	2.30
250	1.78	1.01	3.00	1.22	0.98	2.17	0.87	0.85	2.70
500	0.96	0.77	1.91	0.75	0.67	1.74	0.63	0.60	1.93

It is interesting to observe that the ML estimate of $ë_1$ is usually better than the MHD estimate. The difference in the performance of the two estimators is greater with respect to $ë_2$. An explanation for this might be that, for count data, outliers may occur only at the right tail. As a result, outliers influence more this parameter. In

addition, the covariances of the parameter estimators were, in general, smaller in the case of the MHD method, resulting in the superiority of the MHD method, as judged by the generalised variance ratios reported in Table 5.11.

Table 5.13
Relative MSEs for the ML and MHD methods (correctly specified models)

	p_1=0.2			p_1=0.5			p_1=0.8		
	p_1	$ë_1$	$ë_2$	p_1	$ë_1$	$ë_2$	p_1	$ë_1$	$ë_2$
n				**$ë_1$ =1**		**$ë_2$=2**			
50	3.62	0.98	5.58	1.56	1.00	3.54	1.15	1.13	2.17
100	2.85	0.89	7.57	1.26	0.88	3.33	0.87	0.90	2.13
250	1.96	0.76	4.44	0.88	0.65	3.04	0.55	0.59	2.87
500	1.14	0.58	2.34	0.67	0.50	1.69	0.47	0.48	2.35
				$ë_1$ =1		**$ë_2$=3**			
50	2.14	0.94	3.41	1.09	0.85	1.91	0.81	0.83	1.28
100	2.28	0.92	4.33	0.91	0.81	1.43	0.62	0.67	1.39
250	1.58	0.92	2.48	0.86	0.80	1.21	0.53	0.58	1.08
500	1.23	0.85	1.25	0.81	0.81	0.91	0.50	0.61	0.92
				$ë_1$ =1		**$ë_2$=5**			
50	1.70	1.16	1.60	0.91	1.01	0.78	0.52	0.70	0.71
100	1.33	1.18	0.97	0.93	0.98	0.79	0.66	0.83	0.67
250	1.09	1.06	0.94	0.96	0.99	0.78	0.85	0.94	0.75
500	1.02	0.99	0.93	0.96	0.98	0.87	0.89	0.94	0.80
				$ë_1$ =1		**$ë_2$=8**			
50	0.84	1.01	0.74	0.69	0.98	0.60	0.55	0.90	0.59
100	0.93	1.04	0.81	0.84	0.99	0.70	0.69	0.95	0.57
250	0.98	1.02	0.85	0.93	1.01	0.77	0.90	0.98	0.67
500	0.99	1.01	0.88	0.98	1.01	0.86	0.93	0.99	0.77
				$ë_1$ =2		**$ë_2$=4**			
50	3.23	0.89	4.52	1.37	0.88	2.64	1.12	1.06	1.50
100	2.81	0.84	5.40	1.07	0.77	2.57	0.78	0.76	1.70
250	1.80	0.76	2.94	0.84	0.69	2.05	0.55	0.55	1.78
500	1.20	0.69	1.76	0.77	0.63	1.19	0.41	0.42	1.08
				$ë_1$ =2		**$ë_2$=5**			
50	2.49	0.76	3.16	1.02	0.79	1.61	0.83	0.82	1.02
100	1.97	0.86	2.21	0.85	0.76	1.23	0.59	0.64	1.08
250	1.53	0.95	1.37	0.86	0.83	0.89	0.54	0.62	0.91
500	1.16	0.88	1.07	0.86	0.85	0.90	0.53	0.69	0.76
				$ë_1$ =2.8		**$ë_2$=3.2**			
50	5.32	1.37	7.38	2.52	1.44	6.93	1.69	1.50	4.37
100	5.40	1.16	1.38	1.93	1.30	8.02	1.50	1.51	4.75
250	3.26	1.09	9.21	1.45	0.99	4.62	0.73	0.73	6.94
500	0.92	0.63	3.72	0.54	0.45	2.78	0.38	0.37	3.74

Examining the relative MSEs of the two methods reported in Table 5.13, we can see that, for the parameter $ë_2$ the performance of the ML estimator is inferior. It is the presence of outliers that contributes to this situation. On the other hand, the ML

estimator is more accurate for $\ddot{e}_1$ and in the case of models with well separated components.

Concluding, we can say that the MHD method achieves high efficiency for correctly specified models and it is superior for models with not well separated components.

Lindsay (1994), using second order efficiency arguments, showed that the MHD method is inferior to the ML method. It would be interesting to examine for which sample size this effect occurs and, since for both methods iterative algorithms were used, to examine whether the small difference between the methods is detectable in practice.

5.8.3.2 Contaminated Models

In the sequel, the question of the robustness of the method when the model is not correctly specified is examined. For this purpose, it was assumed that the data come from contaminated 2-finite Poisson mixtures. In particular, it was assumed that an additional component was present at the right tail of the distribution. The probability α associated with this component, (the level of contamination), was allowed to take three values so as to investigate whether the amount of contamination affects the plausibility of the method. Specifically, α was let to take the values 0.01, 0.05, 0.1.

The relative MSE defined as the ratio of the MSE of the ML method to that of the MHD method was used as a measure of robustness. Note that for some of the models considered the notion of contamination is not well defined. The reason is that the level of contamination is very high, relative to the entire sample, with the result that it is not clear with respect to what parameter the mean squared error should be calculated. For example, when the parameter $p_1 = 0.8$ and the contamination level α is high, the probability associated with the second component is very close to the probability assigned to the contaminant, namely $p_2 = 0.18$ and $\alpha = 0.10$.

The relative MSEs for several contaminated models are summarised in by Tables 5.14a through to 5.14d, where $\ddot{e}_3$ is the parameter of the Poisson variable which is assumed to contaminate the 2-finite Poisson model.

Table 5.14a
Relative MSEs based on 1000 replications from a 2-finite Poisson mixture distribution with p_1=p, $ë_1$ =1, $ë_2$=3, and contamination á from a Poisson distribution with $ë_3$=7

	á=0.01			á=0.05			á=0.1		
	p_1	$ë_1$	$ë_2$	p_1	$ë_1$	$ë_2$	p_1	$ë_1$	$ë_2$
n				**p=0.2**					
50	2.84	0.98	9.12	2.91	1.26	8.86	2.05	1.39	4.56
100	2.91	1.09	9.36	2.38	1.55	6.08	1.80	1.61	3.03
250	2.57	1.34	10.73	2.00	1.72	3.65	1.37	1.36	1.81
500	1.84	1.25	5.87	1.69	1.58	2.45	1.22	1.21	1.46
				p=0.5					
50	1.17	0.89	3.62	1.34	1.08	4.40	1.29	1.36	3.53
100	1.13	0.95	4.32	1.45	1.29	3.98	1.34	1.46	2.20
250	1.22	1.07	3.45	1.56	1.56	2.40	1.33	1.42	1.61
500	1.35	1.17	2.26	1.47	1.50	1.75	1.22	1.27	1.36
				p=0.8					
50	0.69	0.74	2.55	0.55	0.70	3.08	0.46	0.87	2.17
100	0.55	0.65	3.31	0.45	0.80	2.68	0.52	1.12	1.78
250	0.54	0.71	3.34	0.92	1.29	1.89	0.95	1.29	1.41
500	0.80	1.03	2.54	1.26	1.40	1.55	1.05	1.23	1.24

Table 5.14b
Relative MSEs based on 1000 replications from a 2-finite Poisson mixture distribution with p_1=p, $ë_1$ =1, $ë_2$=3, and contamination á from a Poisson distribution with $ë_3$=12

	á=0.01			á=0.05			á=0.1		
	p_1	$ë_1$	$ë_2$	p_1	$ë_1$	$ë_2$	p_1	$ë_1$	$ë_2$
n				**p=0.2**					
50	5.58	1.31	44.50	4.96	2.12	16.04	2.76	2.00	5.50
100	6.71	1.85	74.63	3.06	2.48	7.07	1.70	1.70	2.72
250	7.04	3.32	51.94	1.84	1.87	3.01	1.06	1.12	1.32
500	7.98	5.33	48.85	1.29	1.32	1.70	0.99	1.02	1.12
				p=0.5					
50	1.59	1.12	19.43	2.32	1.83	13.85	1.64	1.80	4.33
100	2.03	1.52	43.09	2.25	2.43	6.58	1.35	1.72	2.28
250	3.56	3.11	42.44	1.88	2.19	3.15	1.03	1.18	1.31
500	5.69	5.32	48.36	1.34	1.51	1.85	0.97	1.05	1.11
				p=0.8					
50	0.65	0.76	11.59	0.32	0.73	6.69	0.22	0.96	2.74
100	0.49	0.63	14.63	0.39	0.93	3.28	0.32	1.08	1.61
250	0.67	1.18	18.28	0.86	1.40	1.73	0.70	1.18	1.15
500	1.62	2.88	16.43	0.93	1.23	1.27	0.82	1.12	1.08

Table 5.14c
Relative MSEs based on 1000 replications from a 2-finite Poisson mixture distribution with p_1=p, $ë_1$ =1, $ë_2$=5, and contamination á from a Poisson distribution with $ë_3$=12

	á=0.01			á=0.05			á=0.1		
	p_1	$ë_1$	$ë_2$	p_1	$ë_1$	$ë_2$	p_1	$ë_1$	$ë_2$
n				**p=0.2**					
50	2.81	1.56	8.57	4.01	2.70	11.75	3.10	2.78	7.19
100	2.52	1.74	5.62	4.50	3.24	10.43	3.73	3.36	5.84
250	1.69	1.40	2.05	5.86	4.24	7.60	3.27	2.89	4.04
500	1.40	1.34	1.96	6.69	4.78	6.35	3.17	2.82	3.32
				p=0.5					
50	1.00	1.22	1.94	1.33	2.20	5.48	1.35	2.49	4.45
100	1.03	1.20	1.54	1.62	2.35	4.70	1.66	2.50	3.43
250	1.09	1.19	1.86	1.93	2.58	3.77	1.87	2.45	2.57
500	1.21	1.26	2.24	2.20	2.64	3.08	1.91	2.26	2.24
				p=0.8					
50	0.51	0.88	2.03	0.39	1.02	2.73	0.47	1.48	2.14
100	0.62	0.95	2.03	0.54	1.50	3.03	0.64	1.60	1.90
250	0.97	1.24	2.76	0.95	1.96	2.54	0.86	1.50	1.49
500	1.20	1.47	3.40	1.09	1.83	2.09	1.00	1.35	1.29

Table 5.14d
Relative MSEs based on 1000 replications from a 2-finite Poisson mixture distribution with p_1=p, $ë_1$ =2, $ë_2$=5, and contamination á from a Poisson distribution with $ë_3$=10

	á=0.01			á=0.05			á=0.1		
	p_1	$ë_1$	$ë_2$	p_1	$ë_1$	$ë_2$	p_1	$ë_1$	$ë_2$
n				**p=0.2**					
50	2.91	0.82	6.52	2.93	1.10	9.46	2.37	1.27	5.07
100	2.54	1.01	6.33	2.21	1.46	4.90	1.84	1.54	3.22
250	2.29	1.27	4.47	2.39	1.89	4.28	1.59	1.49	2.09
500	1.60	1.12	2.00	1.99	1.75	2.63	1.41	1.35	1.64
				p=0.5					
50	1.07	0.84	3.45	1.16	1.00	4.37	1.17	1.24	3.20
100	1.03	0.91	3.22	1.47	1.34	3.80	1.41	1.54	2.52
250	1.19	1.04	2.06	1.68	1.66	2.35	1.45	1.52	1.74
500	1.25	1.19	1.95	1.62	1.62	1.89	1.30	1.35	1.44
				p=0.8					
50	0.64	0.67	1.98	0.40	0.55	2.59	0.38	0.83	2.04
100	0.51	0.56	2.75	0.43	0.76	2.50	0.48	1.12	1.81
250	0.57	0.79	2.55	0.84	1.32	2.02	0.86	1.30	1.44
500	0.78	1.08	2.54	1.22	1.41	1.59	1.02	1.24	1.25

Tables 5.14a-5.14d show that the MHD estimator is more robust when the incorrect model is hypothesised, particularly when the sample size is small, the mixing proportion is small and the contaminant is far from the other components. We note again the same behaviour with respect to the parameters, namely that, for $\ddot{e}_2$, the MHD estimator is almost always more robust while, for $\ddot{e}_1$, it is less robust depending on the mixing proportion. In other words, a sort of dependence of the robustness of the method on the mixing proportion is again manifested. Careful examination of the results reveals some points that might give some explanation for this fact. For some samples with large p_1 the other components happened to be represented in the sample with a few observations not far from the origin. As a result, the other components were confounded with the first component yielding an MHD estimate of p_1 that was close to 1. An indication supporting this observation is the increased relative MSEs when the sample sizes were increased.

Concluding, we can say that the MHD method for finite Poisson mixtures is appealing compared to the ML method with respect to both efficiency and robustness. The results support that the MHD method is almost fully efficient when the model is correctly specified, but it possesses the desired robustness property to give robust results when the model is incorrectly specified. Its resistance against outlier observations prevents the MHD method from yielding inconsistent results due to outliers as in the case of the ML method.

5.9 Inferential Procedures for Finite Poisson Mixtures Based on Minimum Hellinger Distance Methods

The properties of its estimators makes the MHD method an interesting tool for other inferential procedures. Any difficulties arising because of the complexity of deriving the MHD estimates has been removed with the HELMIX algorithm. Thus, the use of the MHD method can be extended to cover fields where the domination of ML methods is almost complete.

Some of these procedures will be presented. However it should be emphasised that a lot of the procedures that are based on the ML method (and thus suffering from its problems) can also be based on the MHD method without necessarily increasing

the required effort. Three interesting problems are discussed in this thesis in the direction of using the MHD alternative to the ML as the inferential basis:

- Semiparametric-MHD estimation for finite Poisson mixtures with an unknown number of components.
- Diagnostics based on producing graphs indicative of whether the chosen model is correct.
- Hypothesis testing for finite mixtures using the Hellinger deviance.

The Hellinger deviance test for finite mixtures is presented later in chapter 6.

5.9.1 Semiparametric Minimum Hellinger Distance Estimation

A common problem in applying mixture models is the fact that in many situations the number of components is unknown, as already mentioned in the discussion section of chapter 3. Then, one can proceed either by assuming the number of components to be fixed, say k, and thus search over all the mixing distributions with k-support points, or by letting k to take any integer value. The latter case is the semiparametric case, in which the researcher aims at minimizing an appropriate function over all the mixing distributions with finite support. Semiparametric ML estimation for finite Poisson mixtures was described in chapter 3. In this section the problem of semiparametric MHD estimation for finite Poisson mixtures is examined. The case with known k has been treated in the previous sections.

Assume that d(x) denotes the observed proportion of the value x from a sample of size n and $f_P(x)$ denotes the probability under the assumed model that the random variable X takes the value x, where P denotes the mixing distribution. The mixing distribution P assigns positive probabilities p_i at the points $ë_i$, for i=1, . . ,k and has a parameter vector denoted by **è**. Our aim is to minimize the Hellinger Distance D given by

$$D(d, f_P) = \sum_{x=0}^{\infty} \left[\sqrt{d(x)} - \sqrt{f_P(x)} \right]^2 = \tag{5.25a}$$

$$= 2 - 2\sum_{x=0}^{\infty} \sqrt{d(x) f_P(x)} \quad . \tag{5.25b}$$

Formulae (5.25) are similar to formulae (5.7). The only difference is that the vector of parameters **è** has been replaced by the mixing distribution P. Formula (5.25b) reveals that minimization of the Hellinger distance is equivalent to maximization of the function $\phi(P)$ defined as

$$\varphi(P)=\sum_{x=0}^{\infty}\sqrt{d(x)f_P(x)} \quad . \tag{5.26}$$

This representation reveals in turn some interesting intrinsic properties of the MHD estimation procedure and allows for comparing it to the likelihood method, where maximization of the function

$$L(P)=\sum_{x=0}^{\infty}d(x)\ln\left(f_P(x)\right) \tag{5.27}$$

is required. Formula (5.27) defines the loglikelihood as a function of the mixing distribution, instead of a function of the vector **è** , as in (5.23).

The case of semiparametric ML method for mixture models has been treated by several authors (Simar, 1976, Laird, 1978, Lindsay, 1983a,b, Lesperance and Kaldbfleisch, 1991, Bohning, 1995, among others) and a description of the relative procedures has already been given in chapter 3.

Lindsay (1983a,b) gave the *General Ìixture Maximum Likelihood Theorem* which provides sufficient and necessary conditions for ML estimation in mixture models. This is our Theorem 3.3, in chapter 3.

A similar general theorem for MHD estimation is given in the sequel.

Whittle (1973) derived a theorem for designs in regression problems. For a linear model one wants to know the values of the predictor on which observations have to be taken in order to maximize an optimality function. For example, this function can be the likelihood of the observations or the negative of the variance of the model or several other functions. Finding the best points requires determining the probability measure which assigns positive probability to specific support points of the predictors. This is analogous to our problem where it is also required to determine the points to which the probability measure (i.e. the mixing distribution) must assign positive probabilities.

In the case of MHD estimation the function which has to be maximized is given in (5.26) where P is the mixing distribution. The only condition imposed by

Whittle (1973) is that the function *h* that is being maximized is concave, in the sense of satisfying the sufficient condition

$$\left[\frac{d^2}{de^2}h\left[(1-e)P+eG\right]\right]_{e=0} \le 0\,, \qquad \text{for all measures } P \text{ and } G.$$

It is interesting to note that the Hellinger Distance is concave, in Whittle's sense. To show this it suffices to prove that it satisfies the above concavity condition.

Define $Q=(1-e)P+eG$ and thus $\frac{\partial Q}{\partial e}=G-P$.

Then differentiating the function $\phi(Q)$ given in (5.26) yields

$$\frac{\partial\phi(Q)}{\partial e}=\sum_{x=0}^{\infty}\sqrt{d(x)}\,\frac{f_G(x)-f_P(x)}{2\sqrt{f_Q(x)}}\,,$$

whence

$$\frac{\partial^2\phi(Q)}{\partial e^2}=-\sum_{x=0}^{\infty}\sqrt{d(x)}\,\frac{\left[f_G(x)-f_P(x)\right]^2}{4\left[f_Q(x)\right]^{3/2}}\,.$$

This is clearly negative for all the values of Q which proves the concavity of the Hellinger distance defined in (5.26).

In the sequel, the results obtained by Whittle (1973) for designs are extended so as to apply to distance functions. Define the directional derivative of *ö* at *P* to the direction of an alternative measure *G* as

$$H(P,G)=\lim_{e\to 0}\left[\frac{\phi\left((1-e)P+eG\right)-\phi(P)\right)}{e}\right]$$

For the Hellinger distance this reduces to

$$H(P,G)=\sum_{x=0}^{\infty}\sqrt{d(x)}\left[\frac{f_G(x)-f_P(x)}{\sqrt{f_P(x)}}\right]$$

Of special interest is the case where the measure G is a degenerate distribution at è. In this case the directional derivative is given by

$$H(P,\theta)=\sum_{x=0}^{\infty}\sqrt{d(x)}\left[\frac{f(x|\theta)-f_P(x)}{\sqrt{f_P(x)}}\right]=$$

$$= \sum_{x=0}^{\infty} \sqrt{d(x)} \left[\frac{f(x|\theta)}{\sqrt{f_P(x)}} - \sqrt{f_P(x)} \right] \quad . \tag{5.28}$$

We will refer to this function as the Hellinger -Gradient function to distinguish it from the gradient function defined in (3.6) and used in ML estimation. The Hellinger Gradient function can play an important role in MHD estimation. The following theorem generalizes the results of Whittle and Lindsay:

Theorem 5.2 The mixing distribution $\hat{P}$ is the semiparametric MHD estimate of the mixing distribution if and only if

a) $H(\hat{P},\theta) \leq 0$, for all è not in the support of $\hat{P}$,

b) $H(\hat{P},\theta) = 0$, for all è in the support of $\hat{P}$,

c) $H'(\hat{P},\theta) = 0$, for all è in the support of $\hat{P}$ and

d) $H''(\hat{P},\theta) \leq 0$, for all è in the support of $\hat{P}$

where primes denote differentiation with respect to è.

Proof: The result of the theorem is an immediate consequence of Whittle's theorem for concave functions and the concavity of the Hellinger distance shown above.

This theorem gives the required conditions for the mixing distribution $\hat{P}$ to be the semiparametric MHD estimator. In this case the support size is not restricted. From conditions a) and b) all the support points are maxima of the Hellinger gradient function and, hence, conditions c) and d) also hold.

In some cases the support size is known a priori. In such cases the maximization takes place over all the mixing distributions with the given support size and the following theorem can be shown:

Theorem 5.3 The mixing distribution $\hat{P}$ is the MHD estimate of the mixing distribution with restricted support size if

a) $H(\hat{P},\theta) = 0$ for all *è* in the support of $\hat{P}$ and

b) $H'(\hat{P},\theta) = 0$ for all *è* in the support of $\hat{P}$

Proof: Consider the estimating equations of a k-finite Poisson mixture. These are derived by equating the first derivatives of (5.7) with respect to the parameters to 0 and their final form is as given in (5.8) and (5.9) . From (5.8) we obtain

$$\sum_{x=0}^{\infty}\sqrt{\frac{d(x)}{f_\theta(x)}}f(x,\lambda_j)=\sum_{x=0}^{\infty}\sqrt{\frac{d(x)}{f_\theta(x)}}f(x,\lambda_k) \quad . \tag{5.29}$$

Also, multiplying (5.8) by p_j and adding over $j=1,\ldots,k$ yields

$$\sum_{x=0}^{\infty}\sqrt{\frac{d(x)}{f_\theta(x)}}f_\theta(x)=\sum_{x=0}^{\infty}\sqrt{\frac{d(x)}{f_\theta(x)}}f(x,\lambda_k) \quad . \tag{5.30}$$

Combining (5.29) and (5.30) it follows that

$$\sum_{x=0}^{\infty}\sqrt{\frac{d(x)}{f_\theta(x)}}\left(f(x,\lambda_j)-f_\theta(x)\right)=0 \quad , \qquad \text{for } j=1,\ldots,k.$$

The left hand side of the above equation is the Hellinger Gradient function. Hence, the condition a) has been proved.

To show condition b) observe that for the probability function of the Poisson distribution it holds that $f'(x,\lambda)=f(x-1,\lambda)-f(x,\lambda)$. Then equation (5.9) leads to the result, i.e. the derivative of the Hellinger gradient function is 0 for all j. This completes the proof of the theorem.

The key idea is that in the unrestricted support case the support points are the maxima of the Hellinger gradient function. In the restricted support case the support points are not necessarily maxima. They could be minima or saddle points.

Using the results of Theorems 5.2 and 5.3 a natural procedure for obtaining the semiparametric MHD estimate is the following:

Step 1: Find the solution with k support points.

Step 2: Check, using Theorem 5.2, if the semiparametric MHD estimate has been found. If it has not been found set k:=k+1 and go to step 1.

We may start with k=1, i.e. with the simple Poisson distribution.

Note that algorithms like the VDM algorithm for semiparametric ML estimation (see section 3.7.4) can be used. One, however, should be aware of its shortcomings as described in section 3.7.4.

Specific conditions resulting from Theorem 5.2 are derived in the next chapter where the Hellinger Deviance test for mixtures is introduced.

Note that Bohning and Hoffman (1982) described distance-type estimation methods for probability measures and gave theorems whose results simply require the concavity of these distances. Likelihood and Hellinger methods belong to this class.

5.9.2 The Hellinger Gradient Function as a Diagnostic Tool for the Poisson Distribution

In this section we consider using the plot of the Hellinger Gradient function as a diagnostic tool for detecting if a k-finite mixture is appropriate. Of major concern is to detect if the Poisson distribution (homogeneity model) is an adequate distribution for modeling the data versus a mixture alternative. Lindsay and Roeder (1993) proposed that the plot of the gradient function, given in (3.6), can reveal if the homogeneity model is more appropriate than the inhomogeneity model, i.e. if a simple Poisson distribution is more adequate than a finite Poisson mixture and in general if a k-finite mixture is more adequate than a (k+1)- finite mixture model.

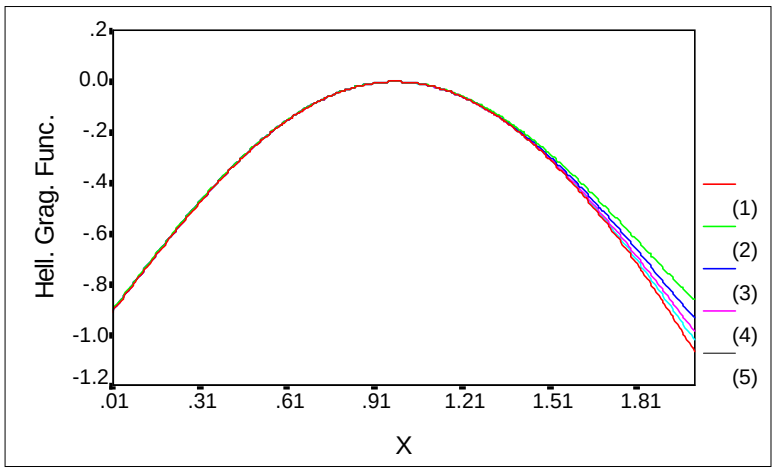
The key criterion is that if the Poisson model is true, the gradient function should be a concave function with maximum attained at the value of the sample mean. Any deviation from this picture reveals departures from the simple Poisson model, keeping in mind that small deviations may have been caused by mere sampling variability. Consider, for example, the plot of the gradient function for the data of example 3.1 which is given in figure 3.4g. Because of the fact that the derived ML estimates are also the semiparametric ML estimates the gradient function has zeroes only at the support points.

A similar approach can be adopted by using the Hellinger gradient function instead of the gradient function. The reason is the robustness of the MHD method, which is lent to the Hellinger gradient function as well.

Of course, the aim of a diagnostic plot is similar to the aim of a detector. It cannot show that something is surely true, but it can reveal if something is clearly false. Diagnostics can simply guide through different choices. So, using the plot of the Hellinger gradient function as a diagnostic tool, the criterion for the validity of the simple Poisson model is the concavity of the function. A non-concave picture it is not proof of non-poissonity, but is simply an indication for this.

Let us now examine more thoroughly this issue. If the Poisson model is true, it follows from Theorem 5.2, that the Hellinger gradient function has a zero only at the value of the MHD estimate of the Poisson parameter, and that it is concave. Thus a plot of the function will provide a picture about the consistence of the assumed Poisson model.

Another important characteristic of the MHD method is its resistance to outliers. This resistance enhances the potential of the Hellinger Gradient function as a diagnostic tool. To see this the gradient function and the Hellinger Gradient function were plotted for a dataset of size 100 in figure 5.7. The vector of observed frequencies (d(0),d(1),d(2),d(3)) were (30, 40, 26, 4). Then a 101-th observation was added at points 5,10,15,20. The effect of this new observation on the gradient functions is depicted in figure 5.7. The new observation influences the gradient function, while the Hellinger gradient function is relatively unchanged and supports the Poisson distribution for all the cases. Note also that when the condition of the concavity of the gradient function is satisfied, the same is true for the Hellinger gradient function. In other words, the Hellinger gradient function remains unaltered while the gradient function changes depending on presence of spurious observations.



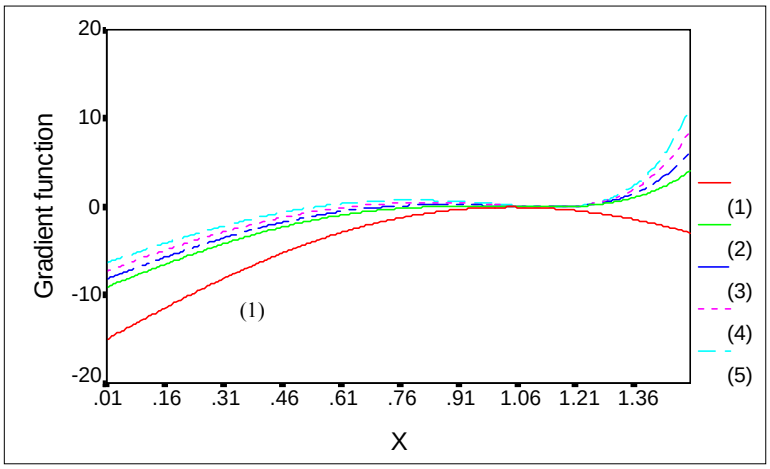


Figure 5.7 The Hellinger gradient function (a) and the gradient function (b) for a dataset generated from a Poisson distribution with mean equal to 1. The sample size was n=100. The vector of observed frequencies were (30, 40, 26, 4). Model (1) refers to these frequencies. Models (2), (3), (4), (5) refer to the cases where an observation was added at 5, 10, 15 and 20 respectively. We can clearly see that these outliers change very much the form of the gradient function, making the Poisson assumption irrelevant, while the Hellinger gradient function is not influenced at all.

It becomes obvious from figure 5.7 that the Hellinger gradient plot can be used as a diagnostic plot. The departure from the Poisson assumption, caused by a simple observation, in the case of the ML method and the robustness of the MHD method to such departures makes the MHD method a preferable diagnostic technique.

Figure 5.8 depicts all the possible combinations of the two gradient functions. The important issue is that the Hellinger gradient verifies the Poisson assumption in all the cases where the simple gradient does so too, but also in cases where the simple gradient function fails to do so because of the presence of outliers.

a b

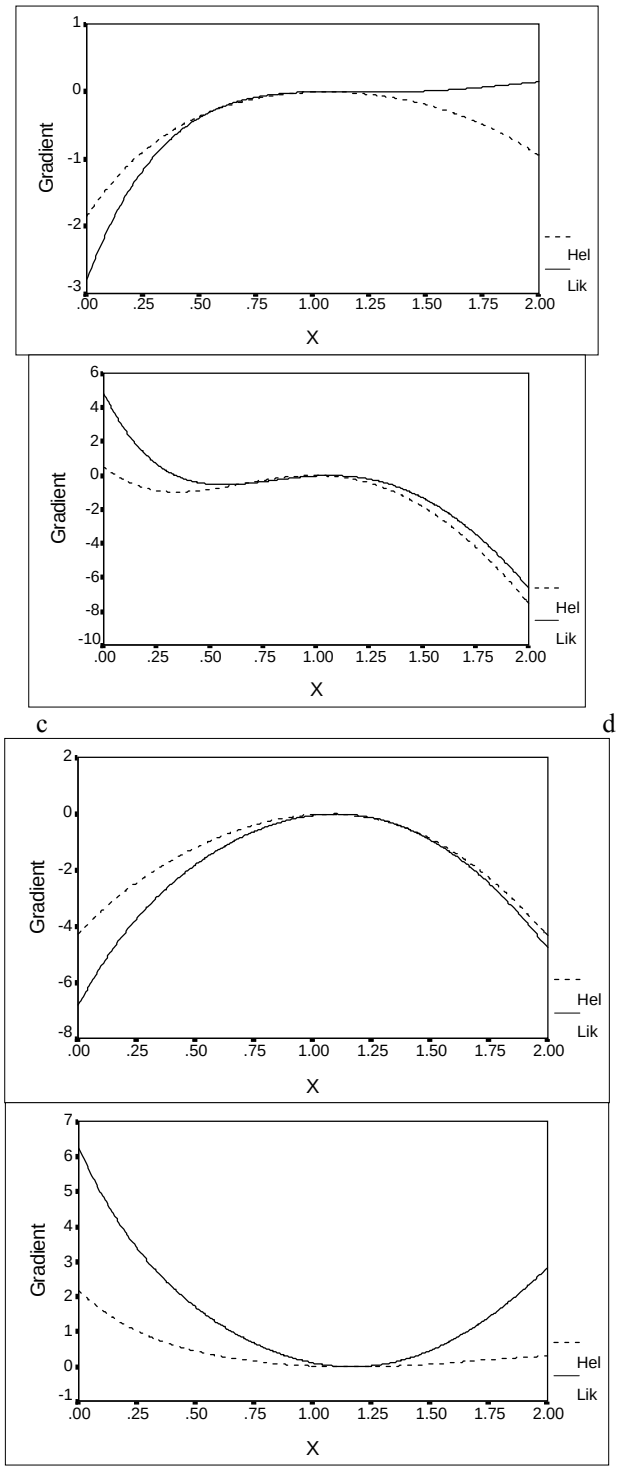


Figure 5.8 The gradient function and the Hellinger Gradient function for samples of size n=100 from a Poisson distribution with parameter 1. The above figures show all the possible cases. Figure c is the case where both the methods support the Poisson distribution, while the likelihood does not support the Poisson distribution for the rest of the cases. Figure a is the case where the two methods disagree.

The Hellinger gradient function can be used for models with more than one component. In each case, the fact that the Hellinger function does not exceed the zero

line, supports the model with k-points of support, while any departure is evidently against this model and a further support point must be added.

Another important issue is the sampling error of the Hellinger gradient function. Lindsay and Roeder (1993) proposed the use of a confidence band, using componentwise asymptotic normality for all the points where the simple gradient function is evaluated. This constructs a zone which, if the Poisson model is true ought to contain a straight line at 0. However, the asymptotic result is rather poor for small sample sizes. A truncated version of the gradient function was also used, because of the unlimited range of the Poisson distribution.

Similar asymptotic results for the gradient function can be derived but their applicability is doubtful because they are hard to derive and a large sample size is needed in order to be meaningful. Clearly, the plot of the gradient function is very useful for a quick check of the Poisson assumption. The concavity of the plot implies that the Poisson model is appropriate, while in any other case we cannot be sure if the non-concavity is due to sampling errors or to systematic departures from the Poisson assumption.

Example 3.1 (continued). Let us go back to example 3.1**,** which refers to the number of crimes committed every month in Greece. We applied the MHD method, and plotted the Hellinger gradient plot (figure 5.9) for the simple Poisson distribution and the 2-finite Poisson mixed distribution. The MHD estimates derived via the HELMIX algorithm were ë=2.157, for the simple Poisson distribution, and $p_1 = 0.641$, $ë_1 = 1.438$ and $ë_2 = 3.5673$, for the 2-finite Poisson mixture model. In figure 5.9 we can see that the simple Poisson model is inappropriate since the Hellinger gradient function does not have a maximum at the MHD estimate, while the 2-finite Poisson mixture is appropriate and has local maxima at the support points.

a b

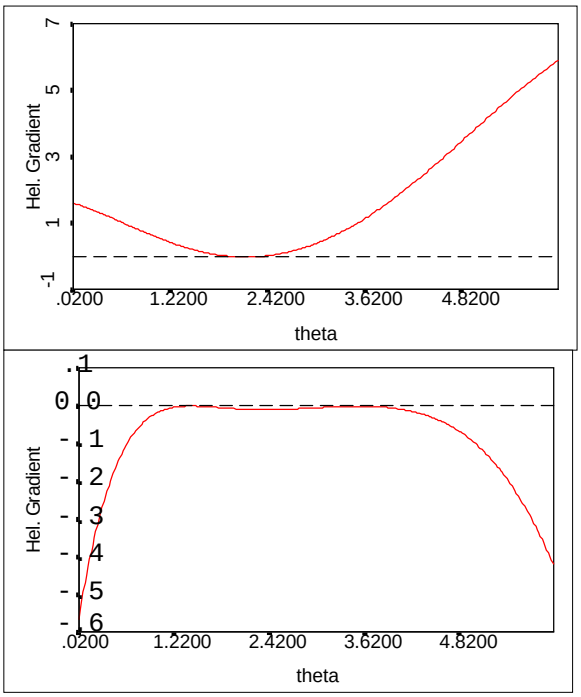


Figure 5.9 The plots of the Hellinger gradient function for the simple Poisson distribution (a) and for the 2-finite Poisson mixture (b), for the data of example 3.1.

5.10 Conclusions

The MHD method for finite Poisson mixtures is both efficient and robust. It is also computationally feasible at a low effort via the HELMIX algorithm. The combination of two potentially useful characteristics makes it an attractive competitor to the ML procedure. It was shown that Hellinger distance based methodologies for diagnostic plotting are very efficient and, at the same time, robust. The latter property is not true for likelihood based inferences, where an outlier may cause inconsistencies. Till now, likelihood based methodologies have attracted almost the entire interest of researchers in the area of Poisson mixtures. However, MHD methodologies seem to be viable (if not preferable) alternatives which can cope with spurious data sets, which makes their use recommendable. Further research would be interesting in order to expand their potential use.

Consider, for example, the problem of likelihood based cluster analysis of rare events given in Symons *et al.* (1983). In such applications, the presence of an outlier can cause problems if the ML estimates are used for obtaining the membership probabilities. Such approaches calculate thresholds which divide the entire line in

segments and then assign observations to these segments. An influenced ML estimate can lead to inconsistent results. A minimum Hellinger based approach can be trustworthy to cope with outliers in such applications.

Extension of the Hellinger based methodology to finite mixtures of other distributions (like the normal or the exponential) is possible. However, it would constitute a tedious task since the MHD estimation is not so clear for continuous models. Cutler and Cordero-Brana (1996) have derived MHD estimators for finite normal mixtures, so MHD based procedures for normal mixtures can be obtained.

On the contrary, extension of Hellinger distance methodology to other discrete distributions is easier. Such an example is the case of finite mixtures of binomial distributions. Putting aside identifiability problems, the HELMIX algorithm can be easily transformed to cover this case too. The key idea is the similarity with the EM algorithm for the ML method.

So far in this thesis the case of MHD estimation was discussed extensively. The Hellinger distance is simply one of the members of the large family of distances a researcher can apply. Inferential reasons make the need for alternative method to the commonly used ML method quite interesting. For example, Albrecht (1980) showed that the familiar chi-square goodness of fit test fails when the ML estimates are used instead of the minimum chi-squared estimates. It is obvious that the derivation of the minimum chi-squared estimates is necessary. Hence, there is a need for an efficient algorithm for deriving these estimates and an appropriate extension of the HELMIX algorithm may provide the answer .

It should be noted that there is scope for considering modifications of the MHD method that will improve its performance in terms of efficiency and robustness (see, e.g., the discussion in Lindsay, 1994). This work can serve as a basis for further investigation on minimum distance methods.

Chapter 6

Robust Testing for Finite Poisson Mixture Models via the Hellinger Deviance Test

6.1 Introduction

The Poisson distribution plays a prominent role in discrete data analysis. It is widely accepted that if the data come from a population at random, the Poisson distribution can very well describe this population. In general, the good fit of the Poisson distribution can be regarded as verifying the assumption that only chance governs the situation under consideration. Good fit of the Poisson distribution is also regarded as a strong evidence for the homogeneity assumption concerning the population under investigation. In this direction, it is very important to be able to test the Poisson hypothesis, i.e. to test the H_0: the data come from a Poisson distribution, against various alternative hypotheses.

Various test criteria have been developed and applied to test this hypothesis. We will focus our attention on the likelihood ratio test (LRT hereafter) as this has become a standard technique for testing a Poisson distribution versus a finite Poisson mixture. An alternative test statistic based on the Hellinger distance is, also, presented. This utilises the robustness properties of the MHD method for testing a hypothesis. An extensive comparison of these two procedures is made.

6.2 The Likelihood Ratio Test for Mixture Models

The LRT is a widely used testing procedure for testing nested hypotheses. The test statistic is calculated as

$$LRT = 2\left[L_1 - L_0\right] \quad , \qquad (6.1)$$

where L_i, i=0,1, is the maximised loglikelihood under the model in hypothesis H_i, i=0,1. Under some regularity conditions, this statistic follows asymptotically a $\div^2$ distribution with degrees of freedom equal to the difference in the numbers of parameters between the two models (Wilks, 1938).

Suppose that we want to test the hypothesis

H_o : The data come from a Poisson distribution ,

against the hypothesis

H_1 : The data come from a 2-finite mixture of Poisson distributions

From (3.2), for k = 2, we can see that the 2-finite mixture leads to a simple Poisson distribution if $p_1 = 0$ *or* $p_1 = 1$. (The case $\lambda_1 = \lambda_2$ has been excluded since the λ_i are in ascending order, a necessary condition for the identifiability of a finite Poisson mixture). So, the set of hypotheses to be tested can be written as:

H_0 : $p_1 = 0$ *or* $p_1 = 1$, against

H_1: $p_1 \in (0,1)$.

Titterington *et al.* (1985, p. 156) showed that the test is equivalent to testing for the number of components in the mixture, i.e. equivalent to testing

H_0 : k=1 against

H_1 : k=2.

Therefore, the problem reduces to one of testing for the number of components (or clusters) in the mixture. This is a very interesting problem in cluster analysis, that has remained unsolved despite the numerous attempts towards its disentanglement. For further information concerning mixture models in clustering the reader is referred to the book of McLachlan and Basford (1988).

A natural testing procedure for such hypotheses would employ a LRT statistic. However, as already mentioned, carrying out this test for mixture models presents some difficulties. The reason for this is that the value of p_1 under the null hypothesis lies on the boundary of the parameter space and hence the regularity conditions fail (see, e.g. , Self and Liang, 1987).

It was shown that if the model is incorrect, the ML estimates are inconsistent. Feng and McCullogh (1996) showed that, even in this case, the likelihood is consistent and thus the test procedure can be applied.

Many attempts have been made in the literature to determine the asymptotic distribution of the test statistic. Titterington *et al.* (1985) showed that the asymptotic distribution of the test statistic is a mixture of a distribution degenerate at 0 and a $\div^2$ distribution with one degree of freedom, in equal mixing proportions. The distribution of the test statistic is the same as the distribution of the random variable Y defined as: $Y = \left(\max[0, X]\right)^2$ where X is a standard normal variate.

Self and Liang (1987) verified this result, while Bohning *et al.* (1994) gave a geometrical representation for the failure of the regularity conditions and verified this result for mixtures of certain members of the exponential family. Vouong (1989) showed the same asymptotic result for the distribution of the LRT in the more general context of model selection and non-nested hypothesis testing. As Lindsay (1983b) showed, the geometrical interpretation of general mixture models allows for constructing a hyperplane where the ML solution is restricted. Adding one more support point outside this hyperplane, the likelihood does not increase and hence the LRT is 0.

In order to avoid the problem of the unknown form of the asymptotic distribution, two main avenues have been proposed. The first one utilises bootstrap methods for constructing the null distribution of the test statistic, by sampling from the distribution in the null hypothesis. This approach has been used by Symons *et al.* (1983), McLachlan (1987), Goffinet and Loisel (1992), Mendell *et al.* (1991,1993), Thode *et al.* (1988), McLachlan *et al.* (1982), McLachlan and Jones (1995), Atwood *et al.* (1996), among others. These simulation experiments verified that the null distribution departs from a $\div^2$ distribution.

The second strategy tries to transform the test statistic or to use a different approximate distribution. One of the initial attempts on this problem was given by Wolfe (1971). He suggested, after a small scale simulation study, that one may use the statistic $2c\left[L_1 - L_0\right]$, where c is the correction factor. This is obtained as

$$c = \left(n - 1 - p - k_2/2\right)/n,$$

where n is the sample size, p is the dimension of the problem (p=1 for the univariate Poisson mixture) and k_2 is the number of components in the alternative hypothesis (2k-1 for a k-finite Poisson mixture). According to Wolfe (1971), this statistics follows a chi-square distribution with degrees of freedom twice the difference in the

number of parameters in the two hypotheses, not including the mixing proportions. Everitt (1981) showed that this approximation works well for small sample sizes in the case of testing a normal distribution against a 2- finite normal mixture. The main disadvantages of such an approach are:

- the chi-square, corrected or not, cannot take into account the possibly high proportion of 0's.
- the transformations are rather empirical without any stable theoretical justification
- the transformations do not work well, in general.

For a detailed description and critique of all of these methods see McLachlan and Basford (1988, pp 22).

In a series of papers, Feng and McCullogh (1992,1994,1996) discussed another approach to overcome this problem. Their proposal was to extend the parameter space so that the parameters, which were on the boundary, to be included in the interior of the parameter space. Then, the likelihood can be maximised in the extended parameter space. One can return to the true parameter space by relating the unrestricted estimator with the restricted parameter space. This estimator is consistent and asymptotically normally distributed (Feng and McCullogh, 1992). Recently, Feng and McCullogh (1996) proposed the use of this method for likelihood testing using estimates for the critical values of the test statistic derived via a bootstrap method.

Aitkin *et al.* (1981) have expressed reservations about the adequacy of such an approximation for the null distribution of the test statistic and they have outlined a solution to the general problem, essentially using a bootstrap approach, in a latent model application. Their bootstrap approach was quite elementary since they based their results on only 19 simulations without examining the performance of the method. However, their approach may be considered as a pioneer approach, adopted by many authors in the subsequent years. A more detailed examination for the distribution, based on simulations, can be found in Thode *et al.* (1988) and Mendell *et al.* (1991).

Recently, Berdai and Garel (1996) and Garel (1998) treated theoretically the normal mixture case, showing that the LRT can be considered as resulting from a Gaussian process. Using this approach, they derived the distribution of the test statistic and they proposed tabulation of its values for the case of normal mixtures.

Another method proposed is the quasi-Bayesian method of Aitkin and Rubin (1985). In order to avoid values for the parameters on the boundary of the parameter

space, they proposed the use of a prior distribution for the mixing proportions, which leads to estimates in the interior of the parameter space whence the standard asymptotic result applies. This result was strongly criticised by Quinn *et al.* (1987). They showed that the gain from such an approach is negligible. Too much computational effort is required to attain accurate results, while the asymptotic result still does not hold.

Two different approaches have been proposed by Chen (1994) and Chen and Cheng (1994,1997). Putting aside the traditional LRT, Chen (1994) derived another generalised likelihood ratio statistic, by partitioning the entire sample into two subsamples and working with them. Chen and Cheng (1994,1997) used a different idea, based on the asymptotic result that half of the times the test statistic is equal to 0. The disadvantage of this approach is that it is based on the asymptotic behaviour of the test statistic. So, with small samples which are common in practice, its performance is not known.

Much, if not all, of the interest has been focused on testing for one component against two components. No attempts seem to have been made for more general cases, apart from a paper by Izenman and Sommer (1988) where a sequential LRT procedure is employed. However, the $\div^2$ distribution is erroneously used as the asymptotic test statistic distribution since, again, the regularity conditions fail to hold. So, testing such hypotheses with the LRT lacks soundness. McLachlan (1992) reviews the use of the LRT for mixture models in the context of discriminant analysis. Soromenho (1994) compared, via simulation, some approaches for determining the optimal number of components in 2-finite mixtures of normal distributions. For a broad review of the problem, from a Bayesian perspective, the reader is referred to Richardson and Green (1997).

6.3 Critical Values for the LRT for Testing the Poisson Distribution Against a 2-Finite Poisson Distribution

As already seen, the null distribution of the LRT statistic is not known and simulation methods are needed for its construction. A simulation experiment was carried out, in order to derive percentile points of the null distribution of the LRT statistic. For several values of the Poisson parameter and for several sample sizes

50000 samples of the given sample size and for the given parameter were simulated, and the value of the LRT statistic was calculated. Then the a-percentile of the distribution was estimated as the [50000a]-th order statistic from the sample of the values of the LRT statistic.

The proportion of zero values for the LRT statistic was also recorded. Figure 6.1 depicts the situation. For selected sample sizes, 10000 samples were simulated from the Poisson distribution with specified parameter values. The values used for the parameter were ë=0.5,..10(0.5),15,20,25,30 and 50. The boxplots of the values of *P(LRT=0),* for all the Poisson distributions used clearly show that as the sample size increases the probability tends to 0.5. Note that in all simulations the probability is greater than 0.5 while, for small sample sizes, the probability is distinctly greater than 0.5, especially for small values of the parameter of the Poisson distribution.

Table 6.1 contains the proportion of zero values for the LRT statistic, for several values of the parameter of the Poisson distribution and several sample sizes. One can see that as the sample size increases, the proportion of values of the LRT statistic tends to 0.5 The same is true when the value of the Poisson parameter increases.

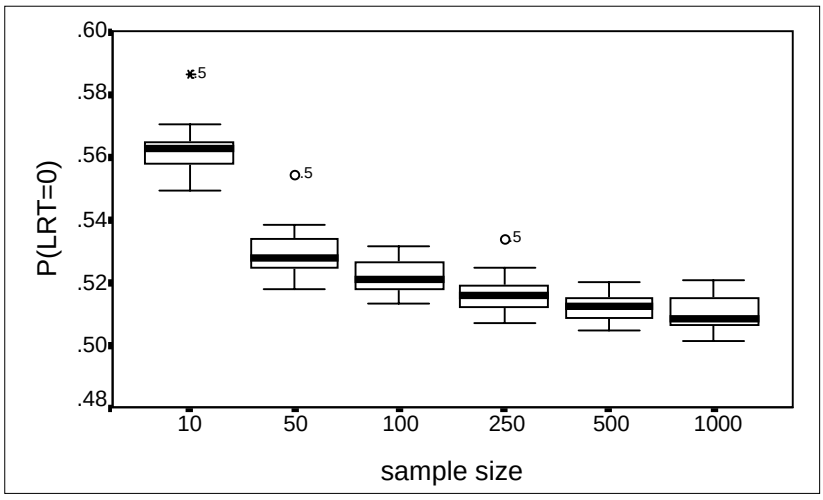


Figure 6.1 The probability that the LRT=0 when we sample from the Poisson distribution with varying parameter value. The boxplots were based on 25 different Poisson means ë=0.5,..10(0.5),15,20,25,30,50. It can be seen that, for all the cases, the probability is greater than 0.5 and, clearly, the proportion of zeroes decreases as the sample size increases. The same is true for the variability of the zero proportion. For small sample sizes, the proportion of zeroes is distinctly larger than 0.5. Note also that the outliers are simulated from the Poisson distribution with mean 0.5.

Table 6.1

The proportion of zero values of the LRT statistic for testing the Poisson distribution against a 2-finite Poisson mixture distribution, for several parameters of the Poisson distribution and sample sizes

	sample size												
ë	10	25	50	75	100	150	200	250	400	500	700	1000	2000
0.3	0.77	0.60	0.58	0.57	0.56	0.56	0.55	0.54	0.53	0.53	0.53	0.52	0.51
0.5	0.70	0.59	0.58	0.57	0.57	0.54	0.55	0.53	0.53	0.53	0.53	0.52	0.51
0.75	0.68	0.60	0.57	0.56	0.55	0.55	0.54	0.53	0.53	0.52	0.52	0.52	0.51
1	0.68	0.59	0.59	0.55	0.56	0.55	0.54	0.53	0.53	0.52	0.52	0.51	0.51
1.5	0.67	0.60	0.57	0.56	0.55	0.53	0.54	0.53	0.53	0.52	0.52	0.51	0.51
2	0.66	0.60	0.57	0.56	0.55	0.54	0.53	0.53	0.52	0.52	0.52	0.51	0.51
2.5	0.66	0.59	0.56	0.56	0.54	0.54	0.53	0.53	0.52	0.52	0.52	0.51	0.51
3	0.65	0.59	0.57	0.55	0.55	0.54	0.53	0.53	0.52	0.52	0.51	0.51	0.51
4	0.65	0.59	0.57	0.55	0.55	0.54	0.53	0.53	0.52	0.52	0.52	0.51	0.51
5	0.65	0.59	0.57	0.55	0.55	0.54	0.53	0.53	0.52	0.52	0.51	0.51	0.51
6	0.65	0.59	0.56	0.55	0.54	0.53	0.53	0.53	0.52	0.52	0.52	0.52	0.51
7	0.65	0.59	0.56	0.55	0.54	0.54	0.53	0.53	0.52	0.52	0.52	0.51	0.51
8	0.65	0.59	0.57	0.55	0.55	0.54	0.53	0.53	0.52	0.52	0.52	0.51	0.51
10	0.65	0.59	0.56	0.55	0.55	0.53	0.53	0.53	0.52	0.52	0.52	0.52	0.51
12	0.65	0.59	0.56	0.55	0.55	0.54	0.53	0.53	0.52	0.52	0.52	0.52	0.51
15	0.65	0.59	0.56	0.55	0.54	0.53	0.53	0.53	0.52	0.52	0.52	0.52	0.51

Tables 6.2-6.4 contain the estimated 90%, 95% and the 99% percentiles . These values can be used as a first check for rejecting the null hypothesis. Unfortunately, as it can be seen in Tables 6.2-6.4, the critical values depend on both the sample size and the Poisson parameter and, hence, a complete tabulation is impossible. So, the researcher may check if the observed value is far from the critical value reported in Tables 6.2-6.4, (using interpolation for values not appearing in the Tables) and thus decide if it is necessary to apply the bootstrap LRT procedure.

The entries of Tables 6.2-6.4 reveal that the LRT statistic is not pivotal. However as the sample size increases the critical values tend to be independent of the parameter of the Poisson distribution. Tables 6.2-6.4 are the first reported tables which tabulate the critical points of the distribution of the LRT statistic for the case of finite Poisson mixtures.

Table 6.2

The estimated 90% percentiles of the null distribution of the LRT statistic for testing the Poisson distribution against a 2-finite Poisson mixture

	sample size												
ë	10	25	50	75	100	150	200	250	400	500	700	1000	2000
0.3	1.072	1.403	1.645	1.759	1.741	1.886	1.945	2.051	2.033	2.018	1.943	1.854	1.817
0.5	1.192	1.677	1.819	1.971	1.978	2.121	2.189	2.199	2.169	2.180	2.069	1.935	1.868
0.75	1.358	1.809	1.966	2.093	2.152	2.217	2.344	2.346	2.322	2.309	2.176	1.978	1.866
1	1.509	1.877	2.099	2.238	2.249	2.334	2.393	2.412	2.425	2.368	2.224	2.017	1.856
1.5	1.591	2.086	2.278	2.323	2.410	2.529	2.602	2.532	2.467	2.449	2.256	1.922	1.864
2	1.879	2.144	2.397	2.474	2.544	2.595	2.617	2.638	2.621	2.428	2.161	1.949	1.873
2.5	1.928	2.313	2.429	2.577	2.549	2.635	2.689	2.681	2.562	2.481	2.117	1.968	1.853
3	1.928	2.306	2.491	2.532	2.675	2.707	2.747	2.695	2.598	2.371	2.123	1.969	1.912
4	2.102	2.309	2.507	2.625	2.648	2.702	2.749	2.716	2.503	2.375	2.109	1.922	1.848
5	2.006	2.391	2.554	2.670	2.668	2.764	2.721	2.655	2.566	2.301	2.101	1.914	1.856
6	1.874	2.466	2.643	2.616	2.647	2.754	2.722	2.719	2.573	2.288	2.052	1.915	1.871
7	1.901	2.234	2.638	2.741	2.755	2.723	2.693	2.654	2.492	2.277	1.998	1.895	1.849
8	1.843	2.274	2.524	2.642	2.680	2.730	2.742	2.613	2.448	2.326	2.014	1.913	1.840
10	1.863	2.279	2.507	2.633	2.668	2.695	2.698	2.657	2.482	2.191	2.034	1.889	1.843
12	1.810	2.284	2.567	2.556	2.662	2.682	2.674	2.613	2.417	2.194	2.001	1.833	1.853
15	1.821	2.261	2.509	2.624	2.683	2.654	2.709	2.599	2.337	2.161	1.925	1.850	1.779

Table 6.3
The estimated 95% percentiles of the null distribution of the LRT statistic for testing the Poisson distribution against a 2-finite Poisson mixture

	sample size												
ë	10	25	50	75	100	150	200	250	400	500	700	1000	2000
0.3	1.854	2.526	2.587	2.911	3.012	3.024	3.039	3.104	3.256	3.310	3.304	3.079	2.945
0.5	1.854	2.832	2.950	3.140	3.228	3.301	3.375	3.378	3.429	3.529	3.472	3.255	3.021
0.75	2.634	2.965	3.198	3.244	3.408	3.490	3.514	3.544	3.604	3.684	3.624	3.288	3.011
1	2.785	3.079	3.255	3.476	3.515	3.580	3.643	3.667	3.649	3.666	3.660	3.428	3.046
1.5	2.862	3.290	3.557	3.535	3.618	3.789	3.872	3.869	3.878	3.897	3.707	3.395	3.040
2	2.989	3.382	3.598	3.749	3.840	3.887	3.960	3.979	3.988	3.962	3.809	3.366	3.043
2.5	3.118	3.536	3.744	3.869	3.826	3.954	3.965	4.056	4.008	4.021	3.739	3.402	2.988
3	3.161	3.544	3.775	3.843	3.978	4.063	4.098	4.084	4.122	3.958	3.749	3.314	3.086
4	3.250	3.694	3.834	3.945	4.042	4.126	4.121	4.160	3.990	3.967	3.692	3.250	3.002
5	3.472	3.577	3.956	4.024	4.089	4.200	4.097	4.075	4.100	3.924	3.672	3.186	3.054
6	3.374	3.897	3.854	4.004	4.101	4.114	4.139	4.235	4.119	3.844	3.539	3.179	3.063
7	3.243	3.735	4.145	4.035	4.080	4.112	4.080	4.095	3.963	3.898	3.572	3.173	2.994
8	3.202	3.564	4.028	4.176	4.229	4.211	4.192	4.054	4.032	3.923	3.582	3.211	3.015
10	3.120	3.638	3.868	3.971	3.988	4.071	4.131	4.151	4.060	3.761	3.541	3.162	2.971
12	3.044	3.652	3.899	3.919	4.058	4.154	4.024	4.130	3.964	3.796	3.502	3.017	2.982
15	3.083	3.535	3.874	4.047	4.077	4.063	4.184	4.103	3.864	3.753	3.435	3.008	2.919

Table 6.4

The estimated 99% percentiles of the null distribution of the LRT statistic for testing the Poisson distribution against a 2-finite Poisson mixture

	sample size												
ë	10	25	50	75	100	150	200	250	400	500	700	1000	2000
0.3	3.720	5.035	5.448	5.855	5.759	5.998	5.974	6.142	6.320	6.180	6.472	6.431	5.725
0.5	4.086	5.663	6.030	5.871	6.267	6.166	6.436	6.422	6.331	6.529	6.534	6.618	6.014
0.75	5.353	5.787	6.056	6.134	6.396	6.479	6.497	6.593	6.531	6.714	6.655	6.783	6.093
1	5.367	5.948	6.136	6.364	6.545	6.524	6.654	6.701	6.725	6.653	6.801	6.630	6.134
1.5	5.642	6.180	6.790	6.486	6.693	6.779	6.870	7.049	7.034	7.043	7.063	6.855	6.217
2	5.816	6.277	6.688	6.722	6.874	6.959	7.045	7.099	7.083	7.231	7.073	6.905	6.061
2.5	6.116	6.495	6.802	6.847	6.831	6.896	7.073	7.287	7.155	7.330	7.018	7.136	6.121
3	6.119	6.587	6.868	6.960	7.076	7.204	7.330	7.248	7.437	7.349	7.265	7.037	5.970
4	6.121	6.778	6.704	7.093	7.160	7.305	7.385	7.523	7.344	7.304	7.313	7.125	5.949
5	6.183	6.979	6.983	7.154	7.193	7.459	7.382	7.437	7.428	7.364	7.346	6.808	5.885
6	6.539	6.620	7.034	7.164	7.175	7.302	7.293	7.523	7.353	7.385	7.228	6.824	6.041
7	6.944	6.729	6.946	7.028	7.394	7.401	7.344	7.235	7.435	7.357	7.287	6.697	5.873
8	6.458	7.266	7.167	7.110	7.309	7.280	7.543	7.206	7.436	7.317	7.080	6.915	5.851
10	6.267	6.900	7.087	7.596	7.327	7.603	7.506	7.407	7.518	7.266	6.979	6.861	5.887
12	6.175	6.819	7.147	7.189	7.311	7.425	7.298	7.366	7.442	7.311	7.124	6.632	5.974
15	6.212	6.749	7.036	7.336	7.249	7.242	7.414	7.334	7.203	7.305	7.065	6.580	5.820

6.4 A Test Based on the Hellinger Distance

6.4.1 The Hellinger Deviance Test (HDT)

The aim of this section is to develop a test based on the Hellinger distance, as a counterpart of the well known LRT. The robustness of the MHD method in the area of estimation motivated the idea of developing a test procedure that may also be robust.

The LRT computes the improvement on the loglikelihood when one more component is added to the model. It would be helpful to derive a test statistic which would measure the improvement of the Hellinger distance if one new component is added. Since the influence of an outlier on this distance is much reduced, compared to that on the likelihood, we expect that this test to be more robust against outliers.

Simpson (1989) proposed the use of Hellinger distance analogues of the LRT for parametric inference. He showed that if the model is correct, the two tests, the one based on the likelihood ratio and that based on the ratio of the Hellinger distances, are asymptotically equivalent. However, the test based on the Hellinger distance is more robust, because of the smaller influence of anomalous data points on the Hellinger distance than on the likelihood. Another test which extends the idea of using the difference in the loglikelihoods between two models to other disparities has been proposed by Read and Cressie (1988) via their general family of power divergent statistics (which contains as special members both the likelihood and the Hellinger distance cases) and the families proposed by Basu and Sarkar (1994) and Lindsay (1994).

The Hellinger Deviance Test (hereafter HDT) statistic is similar to the LRT statistic. The idea in using it is to check if the minimized Hellinger distance for a Poisson mixture is considerably less than the minimized Hellinger distance for a simple Poisson distribution. Hence, the test statistic proposed is given by:

$$HDT = 4n\left[H_0 - H_1\right] \quad , \tag{6.2}$$

where H_i , i= 0 ,1 are the minimized Hellinger distances for the distributions under the two hypotheses (see Simpson, 1989). Under some regularity conditions the HDT statistic follows a $\div^2$ distribution with degrees of freedom equal to the difference in the number of parameters for the two hypotheses (Simpson, 1989). This resembles the well known LRT statistic, discussed in the preceding sections. However, again the

regularity conditions are not satisfied, making the asymptotic result irrelevant. This does not allow the researcher to use the $\div^2$ distribution.

When the model is correct and the sample size large, the estimates obtained by the MHD method and those obtained by the ML method coincide. Simpson (1989) showed that the HDT statistic converges in probability to the LRT statistic. This property indicates that the two tests will have asymptotically the same properties. Figure 6.2 refers to scatterplots of the two tests for 3 different sample sizes. Clearly, as the sample size increases the two test statistics tend to take the same value. However, this convergence is rather slow.

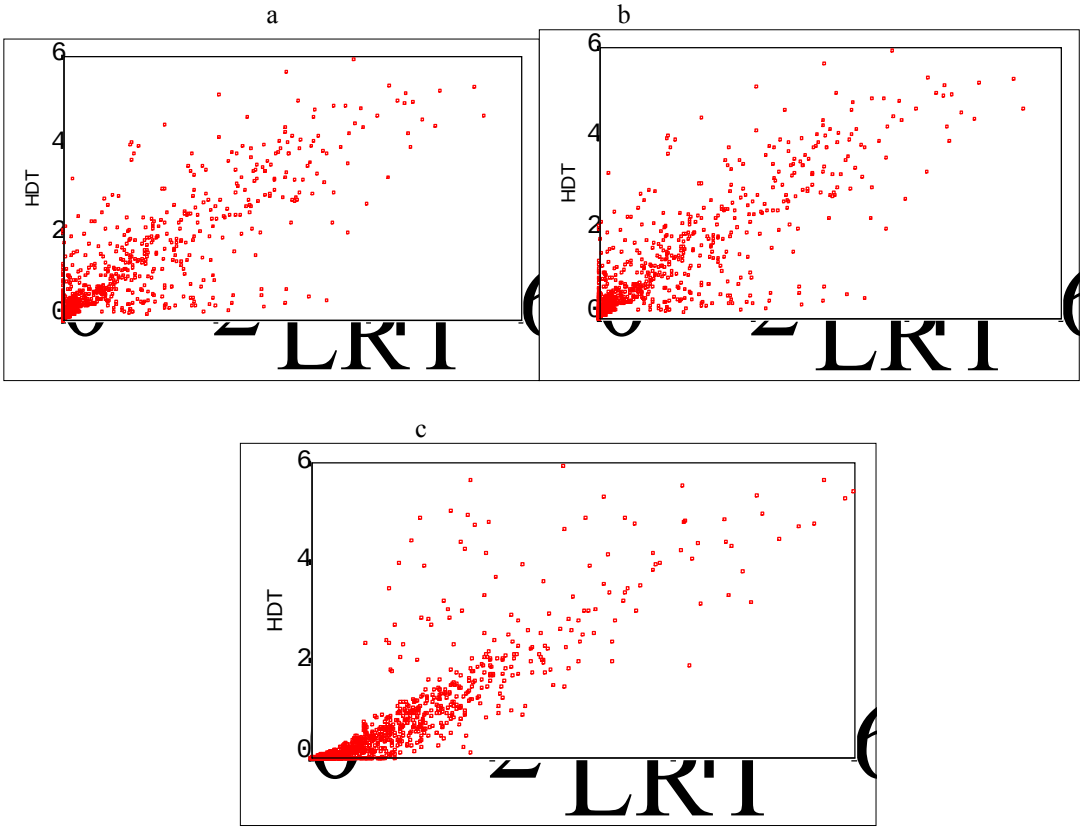


Figure 6.2 Scatterplots of the HDT statistic versus the LRT statistic, for data generated from a Poisson distribution with parameter ë=1. Cases with a 0 value for the LRT statistic have been excluded. Figures a,b,c correspond to different values of n; in particular to the values 50, 250 and 1000, respectively.

Again, the null distribution of the HDT statistic is not known. The ambiguity concerning the distribution of the test statistic limits the usefulness of the test. In order to overcome this shortcoming, we propose the use of a bootstrap test, i.e. a test in which the null distribution of the test statistic is constructed via parametric bootstrap. The test proceeds as follows:

Step 1: Find the MHD estimates of the parameters of the simple Poisson distribution and the 2-finite Poisson mixture, say θ_H and θ_2 respectively, and calculate the HDT statistic, say H_{obs}. The estimates can easily be obtained via the HELMIX algorithm.

Step 2 : Simulate B bootstrap samples of size n, (n is the sample size from the data set) from the Poisson distribution with parameter $\grave{e}_H$ and, for each bootstrap sample, calculate the value of the HDT statistic , say H_j , *j=1, . . . , B.*

Step 3 : Estimate the á-percentile of the distribution of the test statistic by the (100á)-th order statistic from the bootstrap values H_j , *j=1, . . . , B* . Let this percentile be C_a.

If $H_{obs} > C_a$ the null hypothesis that the data come from a Poisson distribution is rejected.

The above scheme can be extended to testing hypothesis of the form:

H_0 : the data come from a k-finite Poisson mixture ,

against

H_1 : the data come from a (k+1)-finite Poisson mixture ,

by replacing $\grave{e}_H$ by $\grave{e}_k$ and $\grave{e}_2$ by $\grave{e}_{k+1}$. In the sequel, we focus our attention on testing the simple Poisson hypothesis. Note that Beran (1988) has shown that bootstrap tests are not inferior to tests based on asymptotic results, recommending the use of bootstrap tests in cases where no exact results are available for the null distribution.

Again, a large proportion of 0 values is present in the distribution of the test statistic. From theorem 5.2 we can see that for certain cases the Hellinger distance cannot be minimized any further by adding a new component. This means that the HDT statistic equals 0. Identifying these cases can substantially reduce the

computational effort required for applying the HDT via the bootstrap procedure, proposed.

The Hellinger gradient function defined in (5.28), combined with the results of theorem 5.2, reveals that we can examine if the semiparametric MHD estimate of the mixing distribution has been obtained, by checking if the support points are the local maxima of the Hellinger gradient function.

From (5.28) the first derivative of the Hellinger Gradient function is given by

$$H'(P,\theta) = \sum_{x=0}^{\infty} \frac{\sqrt{d(x)}}{\sqrt{f_P(x)}} f(x|\theta)\left(\frac{x-\theta}{\theta}\right) \quad .$$

Here, we made use of the fact that, for the Poisson distribution, $f'(x|\theta) = f(x|\theta)\left(\frac{x-\theta}{\theta}\right)$. The second derivative of the Hellinger gradient function is given by

$$H''(P,\theta) = \sum_{x=0}^{\infty} \frac{\sqrt{d(x)}}{\sqrt{f_P(x)}} f(x|\theta)\left[\left(\frac{x-\theta}{\theta}\right)^2 - \frac{x}{\theta^2}\right] \quad . \qquad (6.3)$$

Therefore if the semiparametric MHD estimate has been found for all the support points of this estimate it holds that

$$\sum_{x=0}^{\infty} \frac{\sqrt{d(x)}}{\sqrt{f_P(x)}} f(x|\theta)(x-\theta)^2 < \sum_{x=0}^{\infty} \frac{\sqrt{d(x)}}{\sqrt{f_P(x)}} f(x|\theta)x \quad , \qquad (6.4)$$

Particularly, if k=1, i.e. for testing the simple Poisson distribution against a 2-finite Poisson mixture, condition (6.4) reduces to

$$\sum_{x=0}^{\infty} \sqrt{d(x)f(x|\theta_H)}(x-\theta_H)^2 < \sum_{x=0}^{\infty} x\sqrt{d(x)f(x|\theta_H)} \qquad (6.5)$$

where $\grave{e}_H$ is the MHD estimate for the Poisson distribution.

We can simplify (6.5) further, by noting that the MHD estimate for the simple Poisson distribution satisfies the relation

$$\theta_H = \frac{\sum_{x=0}^{\infty} x\sqrt{d(x)f(x|\theta_H)}}{\sum_{x=0}^{\infty}\sqrt{d(x)f(x|\theta_H)}} \quad . \qquad (6.6)$$

So, (6.5) may be rewritten as:

$$\frac{\sum_{x=0}^{\infty}\sqrt{d(x)f(x|\theta_H)}\left(x-\theta_H\right)^2}{\sum_{x=0}^{\infty}\sqrt{d(x)f(x|\theta_H)}} < \theta_H \quad . \qquad (6.7)$$

Relation (6.7) looks similar to that given above for the ML case where $s^2 < \bar{x}$ (see Lindsay, 1995) The left hand is a weighted second moment around θ_H, while for the ML case, this moment is the variance.

We define the Hellinger ratio HR to be the quantity

$$HR = \frac{1}{\theta_H}\frac{\sum_{x=0}^{\infty}\sqrt{d(x)f(x|\theta_H)}\left(x-\theta_H\right)^2}{\sum_{x=0}^{\infty}\sqrt{d(x)f(x|\theta_H)}} \quad . \qquad (6.8)$$

This ratio is very similar to the variance-to-mean ratio, which is 1 for the Poisson distribution. Figure 6.3 depicts the value of the HDT statistic relative to the Hellinger ratio for different sample sizes. Clearly, as the sample size increases the values of the HDT statistic tend to be concentrated on a curve. Note also that for all the sample sizes, there is a curve which clearly bounds the value of the test statistic given the value of the Hellinger ratio. Figure 6.4 shows the behavior of the LRT statistic with respect to the variance-to-mean ratio. Again, as the sample size increases the values of the LRT can be predicted by the variance-to-mean ratio (index of dispersion). The interesting feature revealed by comparing the figures is that for the HDT, there is a curve which bounds the values of the test statistic. This means that local irregularities of the data set cannot influence very much the value of the test statistic. For example, the inclusion of an outlier in the entire sample cannot influence very much the HDT statistic since its values cannot be lower than these of the bounding curve imposed by the Hellinger Ratio. No similar behavior has been exhibited by the LRT.

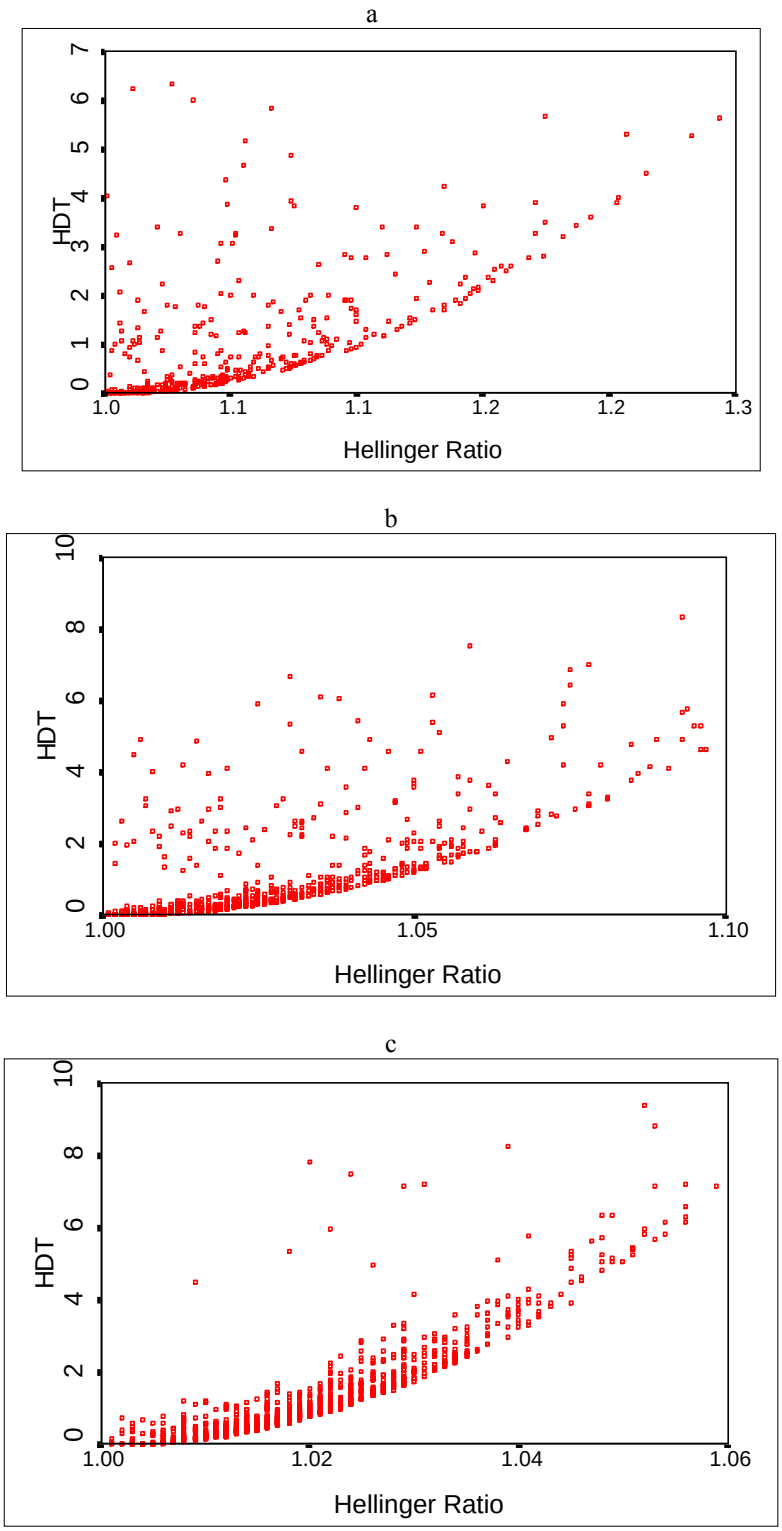


Figure 6.3 Scatterplots of the Hellinger ratio versus the HDT statistic, for data generated from a Poisson distribution with parameter ë=1. Cases with a 0 value for the HDT statistic have been excluded. Figure a, b, c correspond to different values of n (50, 250 and 1000, respectively). We can see that for large sample sizes the value of the HDT statistic is strongly correlated with the Hellinger ratio. Clearly, there is a curve, over which the values of the test statistic are concentrated and there are no values below this curve.

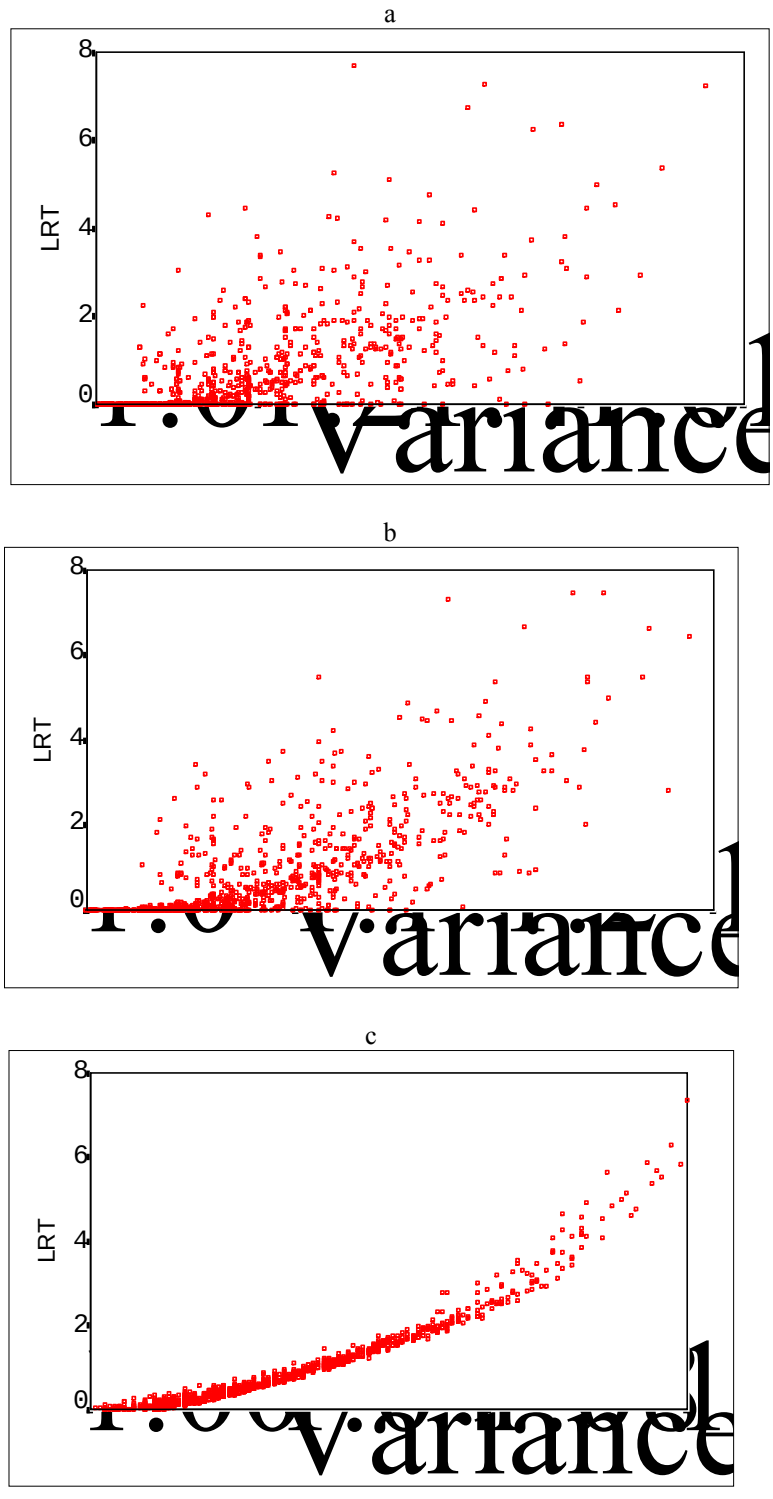


Figure 6.4 Scatterplots of the variance-to-mean ratio versus the LRT statistic, for data generated from a Poisson distribution with parameter ë=1. Cases with a 0 value for the LRT statistic have been excluded. Figures a, b, c correspond to different values of n (50, 250 and 1000, respectively). We can see that for large sample sizes the value of the tests is strongly correlated with the variance-to-mean ratio. Similar plots with the HDT do not reveal such a strong correlation.

The proportion of 0 values for the HDT statistic is much higher than the proportion of 0 values for the LRT statistic. The proportion of 0 values of the HDT statistic was calculated via 50000 simulation for several sample sizes and values of the Poisson parameter and it is depicted in Figures 6.5a-d. The sample sizes used were n=10 (a), n=50 (b), n=100 (c) and n=500 (d). Clearly, the HDT statistic has a 0 value more often than the LRT statistic, for all combinations. It is also clear that the proportion of zeroes for the LRT statistic is relatively stable near 0.5 (but always greater than 0.5) while it increases as the Poisson mean increases for the HDT statistic. An explanation for this is that the Poisson distribution with a large mean gives low probability to a large number of points and the sensitivity of the HDT statistic to such points leads to this phenomenon. However, as the sample size increases the proportion of zeroes decreases.

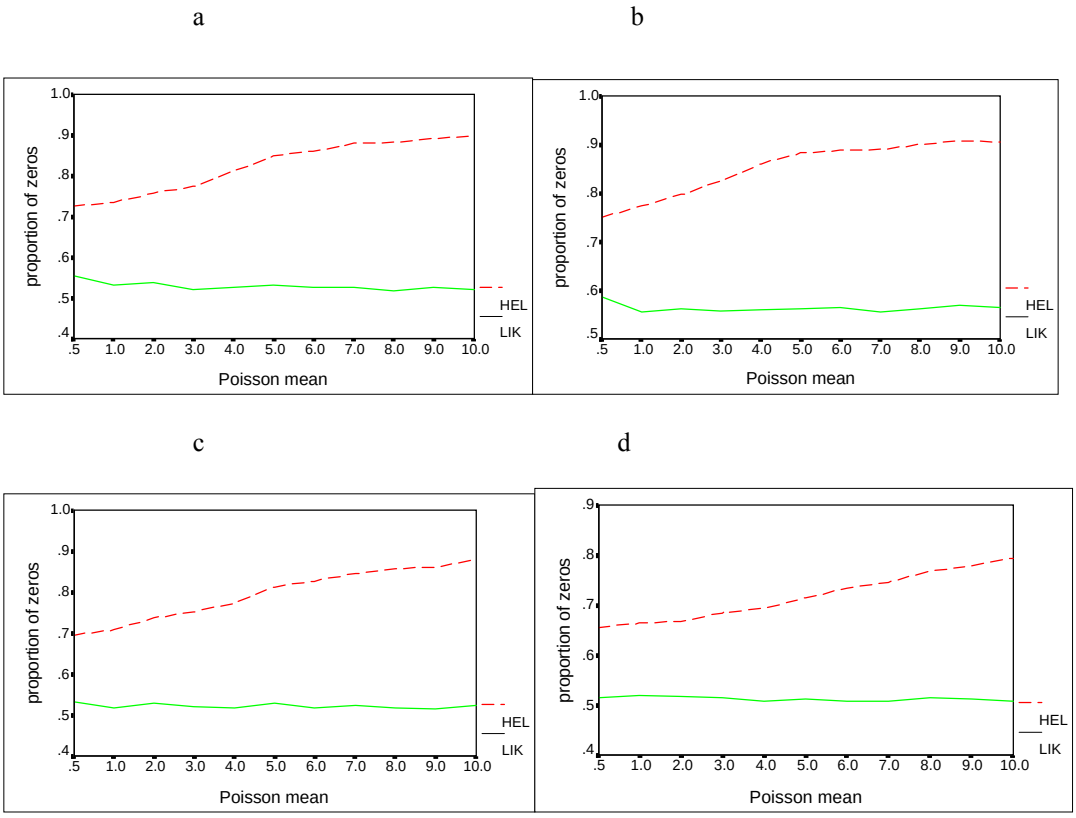


Figure 6.5 The proportion of zeroes calculated via 50000 simulations for each value of the Poisson mean and sample size. The sample sizes used were n=10 (a), n=50 (b), n=100 (c) and n=500 (d).

Another interesting point would be to examine how the occurrence of zeroes of the LRT statistic relates to the occurrence of zeroes of the HDT statistic. We conjecture that there exists a relationship which can be stated as follows.

Conjecture: If the LRT statistic is 0 then the HDT statistic is 0, too. The opposite is not true.

We have not succeeded in proving the above conjecture but we have run more than 100 million simulations with varying configurations of sampling distributions and sample sizes and there has not been any case with a 0 value for the LRT statistic and a nonzero value for the HDT statistic. We hope to be able to report a formal proof of the above conjecture soon.

6.4.2 Critical Values for the HDT Statistic

We carried out a simulation experiment in order to estimate the critical values of the HDT statistic, for several combinations of the values of the Poisson parameter and of the sample size. We applied a slightly different method for finding these critical values. We used the following procedure:

For a given value of the parameter of the Poisson distribution ë and of the sample size n, we simulated 10000 samples of size n from a Poisson distribution with mean ë. For each sample we calculated the value of the test statistic, say H_j, *j=1,. . . , 10000*. Then, the 10000 values H_j were ordered, with $H_{(j)}$ denoting the j-th order statistic. Then, the a% critical point was estimated as $H_{(d)}$, where d=[a*10000], ([a] is the integer part of a). We repeated this procedure 50 times. The entries of Tables 6.5-6.7 are the averages of these 50 repetitions. In figure 6.6 we can see the appropriate boxplots for the critical values.

It is clear from Tables 6.5-6.7 that the critical values are not pivotal and they depend on both the sample size and the value of the Poisson parameter. This makes the complete tabulation of the distribution of the HDT statistic impossible. These few critical values are reported mainly as an initial indication. In practice, the researcher should start by calculating the observed value of the test statistic. If this value is much greater than the reported values (using the necessary interpolation for values not reported in the Tables 6.5-6.7), the researcher has strong evidence to reject the null hypothesis. However, if the observed value is close to the reported values then he/she

ought to carry out his/her own bootstrap for the specific value of the parameter value ë.

The use of a large number of bootstrap samples is strongly recommended. To further support this recommendation one can look at figures 6.6 and 6.7. In these figures boxplots based on 1000 and 10000 bootstrap samples for estimating the 95-th quantile are displayed. Clearly, the range is very wide for accurate inference when only 1000 replications were used, strongly supporting the need of a large number of bootstrap samples in order to increase accuracy. Note also that since the proportion of 0 values is near 0.7, if one uses only 1000 bootstrap samples it is expected that about 700 times a 0 value will be reported and thus the estimation of the right tail is not very accurate. It is worth mentioning that we have verified the entries of Tables 6.5-6.7 up to one decimal point, by repeating the method with 50000 replications.

Table 6.5
The estimated 90% percentiles of the null distribution of the HDT statistic for testing the Poisson distribution against a 2-finite Poisson mixture

	sample size					
λ	20	50	100	200	250	500
0.3	0.69	0.76	0.83	1.04	1.07	1.16
0.5	0.72	0.88	1.05	1.17	1.19	1.22
0.75	0.85	0.99	1.14	1.26	1.29	1.29
1	0.92	1.06	1.19	1.33	1.32	1.28
1.5	0.94	1.09	1.25	1.36	1.38	1.28
2	0.89	1.08	1.22	1.34	1.34	1.16
2.5	0.83	1.00	1.17	1.28	1.26	1.12
3	0.74	0.94	1.10	1.18	1.19	1.07
5	0.11	0.35	0.60	0.79	0.80	0.83
7	0.01	0.07	0.22	0.37	0.42	0.56

Table 6.6
The estimated 95% percentiles of the null distribution of the HDT statistic for testing the Poisson distribution against a 2-finite Poisson mixture

	sample size					
λ	20	50	100	200	250	500
0.3	1.33	1.43	1.65	1.89	1.95	2.09
0.5	1.36	1.66	1.85	2.06	2.11	2.23
0.75	1.56	1.85	2.04	2.22	2.25	2.33
1	1.76	1.93	2.13	2.35	2.32	2.42
1.5	1.85	2.07	2.26	2.41	2.46	2.58
2	1.88	2.09	2.28	2.46	2.50	2.48
2.5	1.83	2.03	2.25	2.46	2.46	2.42
3	1.75	1.96	2.20	2.36	2.42	2.39
5	0.94	1.32	1.66	1.93	1.97	1.95
7	0.49	0.69	0.99	1.26	1.34	1.43

Table 6.7
The estimated 99% percentiles of the null distribution of the HDT statistic for testing the Poisson distribution against a 2-finite Poisson mixture

	sample size					
λ	20	50	100	200	250	500
0.3	2.66	3.35	3.67	4.06	4.19	4.52
0.5	3.31	3.84	4.02	4.39	4.45	4.75
0.75	3.80	4.08	4.35	4.62	4.71	4.83
1	4.02	4.29	4.52	4.81	4.81	5.02
1.5	4.40	4.48	4.77	4.97	5.06	5.31
2	4.42	4.60	4.86	5.07	5.11	5.38
2.5	4.39	4.54	4.77	5.10	5.15	5.31
3	4.33	4.46	4.75	5.01	5.09	5.42
5	3.42	3.71	4.18	4.55	4.62	4.86
7	2.54	2.70	3.18	3.75	3.83	4.33

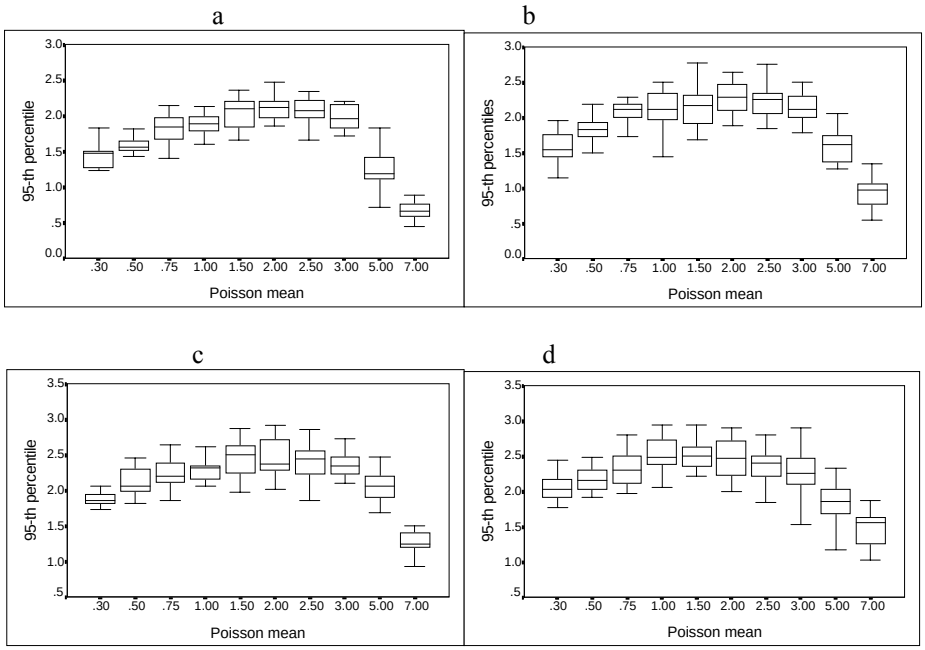


Figure 6.6 Boxplots for the 95% critical value of the HDT statistic, for selected sample sizes (n=50,100,250, 500) respectively for figures a to d. The number of bootstrap samples was set equal to 1000 (B=1000). The boxplots were based on 50 replications for each sample size and Poisson parameter. Clearly, if we use such a small value for B, the variability of the estimated percentile is great, making conclusions based on such replication size questionable.

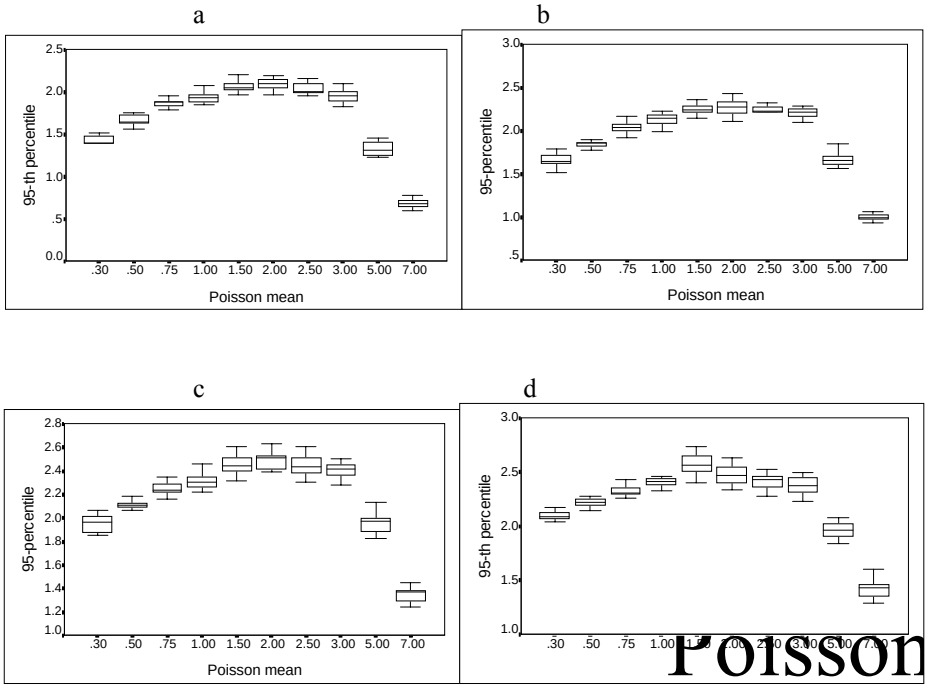


Figure 6.7 Boxplots for the 95% critical value of the HDT statistic, for selected sample sizes (n=50,100,250, 500) respectively for figures a to d. The number of bootstrap samples was set equal to 10000 (B=10000). The boxplots were based on 50 replications for each sample size and Poisson parameter. Note that with the increased replication size the variability has been reduced considerably.

6.5 Power Comparison for the HDT and LRT

In order to examine the performance of the HDT we will examine the power of the test. It was shown in the previous section that the null distribution of the test statistic is not of a known form and simulation was used to construct it. Unfortunately this is also the case for the alternative distribution of the test statistic, which cannot be derived in closed form. To overcome this difficulty we adopted again a simulation based approach. Before going into details we will give the following definition:

Definition 6.1 The empirical power of a test, is the proportion of times the null hypothesis was rejected when the data were generated from the distribution in the alternative hypothesis.

As critical values for the rejection of the null hypothesis, we used the results of the extensive simulation of the previous section. In order to compare the HDT to the LRT we calculated also the power of the LRT. For the LRT statistic the simulated critical values were also used.

Six different alternative distributions were chosen to represent the alternative hypothesis. All these alternatives have the same mean with the mean of the distribution at the null hypothesis. The reason is that, due to the results of section 3.2, the ML estimates under the H_1 satisfy the mean equation and thus in practice the null and the alternative distributions would have the same mean. The 6 alternatives were 2-finite Poisson mixtures with parameter vectors:

A) 0.5, 0.95 ë, 1.05 ë B) 0.5 , 0.5ë, 1.5 ë

C) 0.8, 0.9 ë, 1.4 ë D) 0.8, 0.5ë, 3ë

E) 0.2 , 0.9 ë, 1.025 ë F) 0.2, 0.5ë, 1.125ë

The alternatives were chosen to represent specific kinds of departure form the null distribution. For example, alternative A departs very little from a Poisson distribution. The same is true for alternative F, but now the resulting distribution is more skew. From these alternatives 50000 samples were drawn and the empirical power was calculated. The results are reported in Table 6.8 for values of the Poisson parameter, ë= 1,3,5 and sample sizes n=20, 50, 100, 200, 250, 500.

Table 6.8
The power of the HDT and the LRT

n	ë=1 alternatives											
	A		B		C		D		E		F	
	LRT	HDT	LRT	HDT	LRT	HDT	LRT	HDT	LRT	HDT	LRT	HDT
20	0.053	0.051	0.167	0.182	0.063	0.065	0.452	0.616	0.058	0.054	0.075	0.072
50	0.051	0.047	0.290	0.286	0.066	0.069	0.832	0.902	0.053	0.050	0.088	0.085
100	0.050	0.045	0.453	0.442	0.076	0.071	0.978	0.990	0.050	0.044	0.114	0.102
200	0.051	0.038	0.702	0.684	0.102	0.088	1.000	1.000	0.058	0.047	0.155	0.128
250	0.055	0.040	0.786	0.765	0.109	0.088	1.000	1.000	0.057	0.041	0.180	0.141
500	0.051	0.041	0.960	0.959	0.131	0.119	1.000	1.000	0.052	0.039	0.242	0.208
	ë=3 alternatives											
	A		B		C		D		E		F	
	LRT	HDT	LRT	HDT	LRT	HDT	LRT	HDT	LRT	HDT	LRT	HDT
20	0.054	0.058	0.546	0.581	0.087	0.103	0.544	0.962	0.052	0.055	0.151	0.152
50	0.052	0.050	0.872	0.874	0.110	0.128	0.901	1.000	0.056	0.052	0.239	0.223
100	0.054	0.045	0.992	0.993	0.157	0.164	0.993	1.000	0.055	0.045	0.375	0.331
200	0.053	0.040	1.000	1.000	0.229	0.242	1.000	1.000	0.052	0.036	0.576	0.520
250	0.054	0.039	1.000	1.000	0.273	0.275	1.000	1.000	0.053	0.037	0.662	0.594
500	0.057	0.056	1.000	1.000	0.473	0.538	1.000	1.000	0.057	0.057	0.894	0.873
	ë=5 alternatives											
	A		B		C		D		E		F	
	LRT	HDT	LRT	HDT	LRT	HDT	LRT	HDT	LRT	HDT	LRT	HDT
20	0.051	0.049	0.835	0.862	0.111	0.138	0.299	0.985	0.061	0.059	0.247	0.246
50	0.053	0.054	0.994	0.996	0.175	0.205	0.506	1.000	0.051	0.050	0.451	0.418
100	0.057	0.041	1.000	1.000	0.268	0.292	0.763	1.000	0.060	0.044	0.681	0.621
200	0.061	0.040	1.000	1.000	0.454	0.485	0.943	1.000	0.063	0.044	0.917	0.879
250	0.053	0.043	1.000	1.000	0.510	0.562	0.974	1.000	0.057	0.043	0.951	0.932
500	0.068	0.070	1.000	1.000	0.820	0.866	0.999	1.000	0.072	0.076	0.999	0.999

The entries of Table 6.8 reveal the nice performance of the HDT. The HDT seldom performs worse than the LRT; for several cases (especially for small sample sizes) the difference is substantial. This leads to the conclusion that the HDT is at least as efficient as the LRT is, and its use is a safe guide because of its robustness. This issue is further discussed in the next section. Note that this Table provides evidence for the power of the LRT, too.

6.6 Robustness of the HDT

Assessing the robustness of a test statistic is not a straightforward task. The main problem is that there is not a global definition of the notion of robustness. Usually, we consider a procedure to be robust if a departure from the assumptions does not destroy the performance of the procedure. It is common to consider the two following situations, for robustness studies.

- *Data contamination:* when some observations do not belong to the assumed model and they are included in our data set destroying the underlying assumptions. Such a case is the presence of some outliers at the tails of a distribution and
- *Model deviation:* when the assumed model is not correct and the true model is a little different form the assumed one.

In order to examine the robustness of a test statistic it is useful to discriminate between the two above situations. Consider, for example, the widely used t-test for a normal mean with unknown variance. To apply the above test we assume that the observations are independent and identically distributed following a normal distribution with mean $ì$, the assumed value under the null hypothesis. When a proportion, say p of the observations do not come from the assumed normal distribution but from another Normal distribution with mean , say, $ì_p$, then our data set is said to have been contaminated by these observations. In this case we wish that these few observations do not lead to different conclusions about the data set used.

The notion of model deviation is not much different. If we know that the observations do not come from a normal distribution but from a different distribution, say for example from a t- distribution, we want the test to behave similarly even when the normal assumption is violated.

The above two notions however have a common element. The usual way to describe data contamination is through mixture models. Specifically, we assume that the observations come form a model described as $(1-á)\,P + á\,G$, where P is the true assumed distribution, G is the contaminant which causes the departure from the assumed model and $á$ is the proportion of contaminated values. With this representation of data contamination the two models coincide. However, this representation can help us to examine the effect of a few observations, usually at the

tails of the assumed distribution, when the model deviation implies more general intrinsic departures from the assumed model.

For a goodness of fit test, such as the HDT or the LRT, the situation is more complicated. The reason is that the model deviation approach is misleading. If the data truly come from another distribution, then the test must not select the hypothesized distribution under the H_0. On the other hand, a goodness of fit test which can ignore some spurious observations is very useful in practice. The reason is that for some cases the rejection of a null distribution by a goodness of fit test is caused by quite a few observations. In the sequel we examine only data contamination models.

Robustness of tests has been examined for several tests and from several points of view. Ylvisaker (1977) examined the resistance of a test which is defined as the smaller proportion of observations which can determine the decision ignoring the values of all of the remaining observations. Later, neglecting the acceptance-rejection approach of testing statistical hypothesis, Lambert (1981) proposed the use of the Influence Function (IF) to examine the behavior of statistical tests. He proposed that the IF of the p-value can reveal the robustness properties of test statistic. Knowledge of the form of the null distribution is not always possible. However, this approach can work in cases where the test statistic has a clear, closed form expression even if its null distribution is not known. Hertier and Ronchetti (1994) have shown that the influence curves of both the level and the power of a test are proportional to the influence curves of the estimators used. Later He *et al.* (1990) examined the power breakdown points of test statistics. The power breakdown point is the amount of contamination of each alternative distribution that can carry the test statistic to a null value. See also Lambert (1982) for a qualitative examination of test robustness.

Simpson (1989) and Lindsay (1994) have shown that tests based on the Hellinger distance can be more robust than those based on the likelihood carrying the robustness of the Hellinger estimators.

For our case, the problem that the null distribution is not known and has to be estimated via simulation, prevents us from fully adopting the above mentioned approaches. However, in order to demonstrate the superiority of the HDT relative to LRT, some comparisons are made. The first approach is the use of the IF of the test statistic, while the latter is the simple examination of the performance of the tests when some contamination is present.

The IF defined in (5.10) is a useful tool for examining the robustness of a procedure. From (5.28), one can see that, by definition, both the Hellinger gradient function and the gradient function, are IF for the corresponding distances when a new component is added. Hence, the examination of the gradient functions themselves reveals interesting robustness properties of the two methods, as seen in section 5.10.

It would be interesting to examine the IF for the corresponding distances for the two methods defined in (5.26) and (5.27). It can be seen that the IF for the two distances in (5.26) and (5.27) (the Hellinger distance and the loglikelihood respectively) are given by:

$$IF(z,L,F) = -\sum_{x=0}^{\infty} d(x)\ln f(x) + \ln f(z) \quad , \qquad (6.9)$$

for the likelihood and

$$IF(z,\phi,F) = \frac{1}{2}\left[\sum_{x=0}^{\infty}\sqrt{d(x)f(x)} + \sqrt{f(z)}\right] \quad , \qquad (6.10)$$

for the Hellinger distance.

We used the De L'Hospital rule for deriving the required limit for the IF. The derivation of (6.10) is much complicated but the results stem from the equation

$$\phi\left[(1-t)F + t\Delta_z\right] = \sqrt{1-t}\,\phi(F) + (1-\sqrt{1-t})\sqrt{f(z)}\,.$$

(6.9) and (6.10) represent the influence of a new observation at z, on the likelihood and the Hellinger distance and not on the optimized versions of them. Moreover, the signs of these functions are irrelevant, since the likelihood is known to be negative, while the Hellinger distance is always positive. In both (6.9) and (6.10), the probability function $f(x)$ is calculated using the corresponding estimates.

If z is an outlier, we expect $f(z)$ to be very small, i.e. very close to 0. Since the logarithm near 0 decreases sharper, the IF is also sharper. This indicates that the MHD is not influenced so much by an outlier.

On the other hand, the test statistics associated with the two methods will have IF which, ignoring constants, will depend on $\sqrt{f_1(z)} - \sqrt{f_0(z)}$, for the MHD method and on $\ln\{f_1(z)/f_0(z)\}$, for the ML method, where the subscript under the probability function refers to the distribution used as determined by hypothesis H_i, *i=0,1*.

To see this result, consider the LRT statistic defined in (6.1). Suppose, that we have calculated the LRT when one new observation at z, occurs. Then the IF of the LRT statistic at the point z, will be given as:

$$IF(z, L, F) = \sum_{x=0}^{\infty} d(x) \ln f_0(x) - \sum_{x=0}^{\infty} d(x) \ln f_1(x) + \ln f_1(z) - \ln f_0(z) \ .$$

The first two terms comprise the already calculated LRT, without the added observation, and hence the influence of the new observation is the remainder, i.e. $\ln\{f_1(z) / f_0(z)\}$. Using similar arguments, we can conclude that, for the HDT, the influence is measured by $\sqrt{f_1(z)} - \sqrt{f_0(z)}$.

Two facts support the superiority of the MHD method. The first is that if an outlier is present, the MHD estimates do not differ too much between the two models (see the robustness of the MHD estimate derived in chapter 5) and then we expect an influence close to 0, while for the ML method the change of the estimates causes a positive influence. It is known that a mixed Poisson distribution has thicker tails than the simple Poisson distribution with the same mean (Shaked, 1980). For testing purposes, the means of the two models are assumed to be equal (see section 2.2) and thus the ratio $f_1(z)/f_0(z)$ is greater than 1. Hence, the influence is always positive.

Therefore , an outlier has always a positive effect on the LRT statistic, which always increases if a new outlier observation is added, while the HDT statistic may be stable, or it may increase much less.

Some empirical results support further the above mentioned issue. Suppose now that the functional *T(F)* is the corresponding test statistic for the two methods. Since this statistic does not have a closed form, we are not able to compute the exact IF. An alternative approach is the use of the Empirical Influence Function (EIF). According to Hampel *et al.* (1986, pp. 93), the EIF of the estimator based on any sample is a plot of the values of the estimator, if one more observation (contaminant) is added at the point x.

So, the EIF was used to examine the behavior of the two tests. In order to avoid calculating the EIF for only one sample, 1000 samples of size n=20,100,250,1000 were drawn from a Poisson distribution with parameter *ë=1*. The EIF was calculated if a (n+1)-th observation was added at point x and the averaged influences for all the points x=0,1, ..., 20 were computed. With such an approach,

results due to sampling errors were eliminated so that one can have clearer and more reliable picture about the robustness of the two tests to contamination.

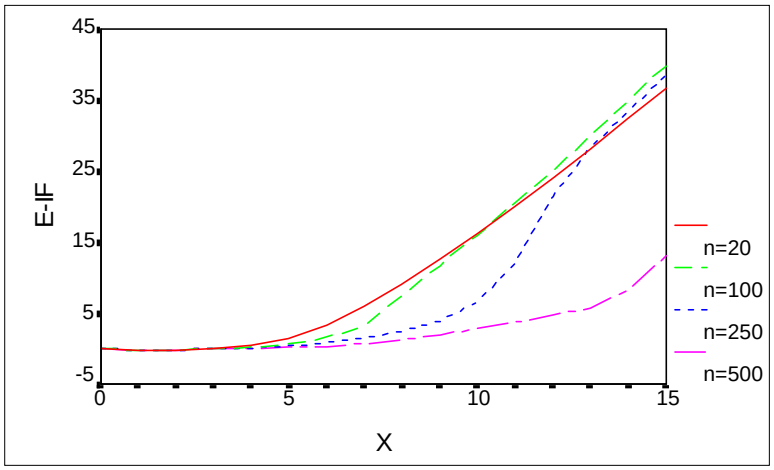


Figure 6.8The E-IF for the LRT statistic, when one more observation is added at point x.

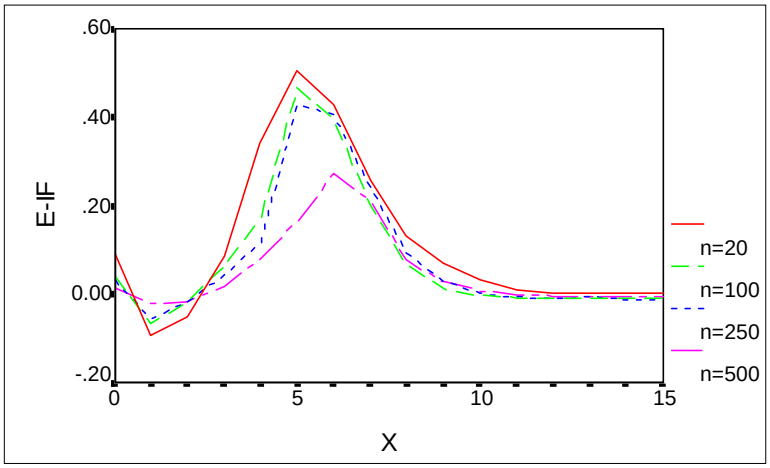


Figure 6.9 The E-IF for the HDT statistic, when one more observation is added at point x

Figures 6.8 and 6.9 depict the EIF, for the LRT and the HDT respectively, when one observation is added at point x. One can clearly see that the LRT can lead to incorrect conclusions when just one observation is an outlier. It is also interesting that this may happen even if the sample size is as large as 500 and the added observation takes a value which is not far from the main body of the data (e.g. x=7). On the contrary, the HDT ignores observations far from the rest. Note also that whatever the point of contamination, the HDT will not reject the hypothesis that the data come from a Poisson distribution, as indicated by the small value of the E-IF. Note the great difference in the scale for the two tests. For the HDT the largest difference is smaller than 0.6, when the difference for the LRT is greater than 5 for an added point at x=5. If we had used the same scale as for the LRT, the HDT would have resembled a straight line!

The interesting point, brought by Figures 6.8 and 6.9 is that a single observation can lead to an incorrect decision based on the LRT. This cannot happen with the HDT.

The behavior of the test when more outliers are present is also examined. In order to do this samples from a contaminated Poisson with mean 1 were taken. The contamination was effected through a degenerate distribution at the point x = 8 and 12 respectively. The proportions of contamination considered were α=0.01, 0.02 (i.e. an outlier at x is drawn with probability a). A robust test ought to cope with such a case of contamination, in the sense that the significance level of the test ought not to increase very much. The significance level was set at 5%. Table 6.9 contains the true significance level when we sampled from the 4 models described above, for both tests. All the entries of the table were based on 10000 simulation samples.

Table 6.9
The calculated significance levels of the HDT and the LRT for contaminated models. The actual level is 5%

models used in the comparison								
sample size	A x=8,a=0.01		B x=12, a=0.01		C x=8, a=0.02		D x=12, a=0.02	
n	HDT	LRT	HDT	LRT	HDT	LRT	HDT	LRT
20	0.054	0.222	0.050	0.224	0.063	0.426	0.050	0.428
100	0.073	0.620	0.050	0.652	0.114	0.907	0.052	0.925
250	0.090	0.756	0.049	0.913	0.134	0.986	0.048	0.998
500	0.094	0.935	0.044	0.986	0.159	1.000	0.050	1.000

From the entries of Table 6.9, it can be seen again that the HDT is far more robust. When the contamination is at x=12, the HDT almost ignores this observation. Note that for n=500 and a=2% we have 10 outliers and the HDT ignores them. On the contrary, the LRT cannot cope with this type of contamination and as the sample size increases it almost surely rejects the null hypothesis.

It should be emphasized again that robustness and power are rather conflicting issues for tests, especially when one wants to examine goodness of fit tests as it is our case. The reason is that we want a sensitive test which can detect departures from the model under the null hypothesis. So, if a test is very sensitive, few observations can destroy its performance. In this sense, it is preferable to find a test which is not so sensitive and it can detect 'faults' which are not caused by the alternative hypothesis,

but from a contamination mechanism. HDT seems to be such a test, combining high power when the data are not contaminated and robustness when the data have been contaminated.

6.7 Conclusions

In this chapter we derived a test procedure based on the Hellinger distance. It was shown that this test is far more robust than the LRT, which is widely used for testing hypotheses in mixture models. Obviously, the HDT can be extended to mixture models of other families, such as normal mixtures, mixtures of the binomial distribution etc. The results showed that the HDT is much more closely related to the Neyman *C(α)* test (see, Lindsay, 1995). This test is simpler as it does not require iterative calculations. However, the nice robustness properties of the HDT makes it a reasonable choice.

This hypothesis testing procedure, that was introduced, completes the MHD based methodologies described in Chapter 5. Inferences based on the Minimum Hellinger Distance can replace likelihood based inferences as the former achieve the dual goal of high efficiency and high robustness. The algorithm HELMIX provides a useful tool for deriving the MHD estimates. No clear advantage of the ML method can be found relative to the MHD method. MHD based methodologies are as efficient and as easy to apply as the likelihood based methodologies, while at the same time they are more robust.

Chapter 7

Determining the Number of Components in a Mixture

7.1 Introduction

Mixture models are widely used to describe inhomogeneous populations, i.e. populations which can be assumed to consist of several subpopulations. In particular, k-finite mixture models may be considered as describing a population consisting of k subpopulations. An interesting problem is to determine the number of subpopulations comprising the entire population, i.e. to determine the number of components in a mixture. The physical interpretation of such a result is of particular interest, since it may give information about the structure of the whole population.

This problem has attracted the attention of the statistical community and it has induced a lot of effort towards its solution. However, no unique solution of this problem is known as yet. In the previous section, testing procedures for testing for a mixture with k components versus a mixture with k+1 components were examined. In this chapter, such tests are employed with the aim of developing a procedure for determining the number of components.

The literature is very sparse of results for testing in a mixture the existence of more than two components. The problem of determining the number of components in a mixture is similar in nature to that of determining the number of clusters in cluster analysis. Since the mixture approach to cluster analysis is widely used, any methodology for determining the number of components in a mixture will constitute a

useful source of techniques for determining the number of clusters in the area of cluster analysis.

This chapter is devoted to this problem, starting with a review of existing methods. A new method for determining the number of components in a mixture is presented and applied to finite Poisson mixtures. This new method is based on the sequential application of the LRT for mixture models discussed in the previous chapter. A simulation experiment for examining the performance of the new method is also given and applications with real data sets are provided.

7.2. A Review of existing methods

Many researchers have proposed methods for determining the number of components in a mixture model. Nevertheless, there does not exist a global procedure for answering this question. Several of the existing methods are reviewed in the sequel.

It should perhaps be noted at this point that very few methods exist for determining the number of components in Poisson mixtures. The majority of the methods have been developed for normal mixtures, mainly because of the wide applicability of normal mixtures in cluster analysis and related fields of statistical methodology, like pattern recognition and discriminant analysis.

Leroux and Putterman (1992) proposed penalised ML methods, for a Markov-dependent Poisson mixture model. This penalised version introduces a penalty term in the likelihood so as to discourage the selection of an excessive number of components. The authors proposed to use two criteria:

the AIC (Akaike Information Criterion) which utilises the function

$$A(m) = L_m - d_m$$

and the BIC (Bayesian Information Criterion) which utilises the function

$$B(m) = L_m - \ln(n)d_m / 2 \ ,$$

where L_m is the maximised loglikelihood for a model with m components, d_m is the number of free parameters in the m-component model, which, for a m-finite Poisson mixture distribution is equal to 2m-1, and n is the sample size.

These criteria select the model with k components, if $A(k)$ ($B(k)$, respectively) is the maximum of the function $A(m)$ ($B(m)$, respectively). These two criteria are widely used in several other model choice problems, such as regression problems etc.

Both of these criteria have certain serious disadvantages.

The AIC is a very conservative criterion. If ℓ is the value of the LRT statistic for choosing between a model with k-1 components and a model with k components, then at every step the criterion selects the model with k components only if

$$\ell = 2\left(L_k - L_{k-1}\right) > 4 \quad .$$

This is so, since the model with k components is selected if $A(k) - A(k-1) > 0$. This implies that

$$A(k) - A(k-1) > 0 \Leftrightarrow L_k - d_k - L_{k-1} + d_{k-1} > 0 \Leftrightarrow$$
$$L_k - L_{k-1} > 2$$

which coincides with the above mentioned condition.

This condition is used whatever the value of k is and for any sample size. In other words, a new component is added only if the loglikelihood is increased by 2 in absolute value. The criterion ignores both the number k and the sample size and this may be quite misleading, since the absolute value of the loglikelihood depends clearly on the sample size. On the other hand, when a large number of components have already been determined for the model, the relative change of the loglikelihood is different when one more component is added.

The BIC takes the sample size into account, but it is in fact more conservative than the AIC. The condition used for accepting the model with k components in favour of a model with k-1 components is

$$\ell = 2\left(L_k - L_{k-1}\right) > 2\ln(n)$$

which clearly depends on the sample size.

This is so, since we select the model with k components if $B(k) - B(k-1) > 0$. This implies that

$$B(k) - B(k-1) > 0 \Leftrightarrow L_k - \frac{\ln(n) d_k}{2} - L_{k-1} + \frac{\ln(n) d_{k-1}}{2} > 0 \Leftrightarrow$$

$$L_k - L_{k-1} > \ln(n)$$

which proves the above mentioned condition.

Again, this criterion ignores the value of k and may thus be misleading. Both criteria favour models with a few components. Note that even for small sample sizes, e.g. n=10, the BIC adds a new component with more difficulty than the AIC as the former reduces to ℓ >4.6 since $\ln(10) \approx 2.3$.

Leroux (1992) proved that, under mild conditions, the estimator of the mixing distribution (finite or not) obtained with the number of components selected using AIC or BIC (or certain other criteria) is consistent and has, in the limit, at least as many components. However, both criteria have some disadvantages. They ignore the number of already involved components. The AIC also ignores the sample size. In practice, both criteria are rather conservative in favour of models with very few components.

Recently, Chen and Kalbfleisch (1996) proposed penalised minimum distance methods for finite mixtures. The innovation of this paper lies in the fact that the penalty function was based on the mixing proportions, so as to discourage components with small mixing proportions. The penalty function for a model with k components was defined to be proportional to $\sum_{j=1}^{k} \ln p_j$. One can see that the contribution of components with small mixing proportion p_j to the value of the penalty function is high. Hence the method discourages the choice of components with small mixing proportions. The authors applied their method to finite Poisson mixtures using the Kolmogorov distance. Henna (1985) proposed another method for selecting the number of components using minimum distance methodology.

Celeux and Diebolt (1985) described the Stochastic EM (SEM) algorithm for mixtures (see section 3.4) . They proposed to use this scheme for estimating both the unknown parameters and the number of components by dropping components with very low mixing proportions. Leonard *et al.* (1994) proposed a similar approach for the EM algorithm. In particular, they proposed dropping components with very small proportions and combining components with parameter very close in value. In

practice this scheme depends on the choice of good initial values. Note that this algorithm often yields the semiparametric ML estimate.

As far as moment estimation is concerned, Lindsay (1989) gave a procedure for selecting the number of components. He based his method on adding a new (k+1)-th support point if the moment problem for a mixing distribution with k support points is solvable (see section 4.3). However, it is known that moment methods are not efficient and, as seen in chapter 4, they often result in a very small number of components.

Fruman and Lindsay (1994a) used the property of mixture models that the variance can be decomposed into two terms, one due to randomness and one due to mixing, as described in section 1.2, for general mixtures, and in section 2.2, for Poisson mixtures (see also Lindsay, 1989). They applied this idea to normal mixtures. This property is known to hold for Poisson mixtures as well. Since the total variance can be decomposed into two parts, the authors suggested adding one component at each step until the improvement in the part of the variance explained by the mixing distribution ceases to increase. Their idea is identical to that used in multiple regression where the coefficient of determination always increases when a new explanatory variable is included in the model, but one tests whether this increase is statistically significant. The main difference is that Fruman and Lindsay (1994a) proposed using a bootstrap test for deciding whether to add one more component based on a moment estimation.

Lindsay and Roeder (1992, 1997) and Roeder (1994) suggested the use of residual diagnostics for determining the number of components in a mixture. The authors showed that smoothed residuals, obtained from the fitted mixed model, provide information about the number of components. In fact, the gradient function used by the authors is the same as the one used to check if the semiparametric ML estimate was obtained.

Windham and Culter (1992) proposed the use of the ratio of Fisher information matrices for selecting the number of components. Their method is termed the Minimum Information Ratio Estimation and Validation (MIREV). The procedure is based on the smallest eigenvalue of the matrix $\mathbf{F}_c^{-1}\mathbf{F}$, referred to as the Minimum Information Ratio (MIR), where $\mathbf{F}_c$ is the information matrix for the classified sample

(in this case the memberships of the observations are known) and **F** is the information matrix for a model with a given number of components. The entries of the matrix $\mathbf{F_c}$ can be derived by using the derived ML estimates in order to calculate the posterior probability for each observation that belongs to a particular component. Then, each observation is classified in the component for which the posterior probability is maximum. In this way a new sample is obtained in which all the observations are classified. This sample is then used for deriving the information matrix $\mathbf{F_c}$.

The information ratio would then measure the proportion of information available without knowing the subpopulation memberships. The procedure can be used sequentially as follows:

- Choose k_1 and k_2 with $2 \leq k_1 \leq k_2$.
- For each k, $k_1 \leq k \leq k_2$, obtain the MIR_k assuming a mixture with k components.
- The value of k for which the MIR_k is largest is the estimate of the number of components.

After selecting the number of components, they proposed a validation technique based on bootstrapping, in order to examine the performance of this estimate. The validation steps are:

- For $m=1, \ldots, M$ obtain a bootstrap sample from the original data and compute from this data set the estimate k_m with the procedure described above.
- Estimate the probability that the maximum MIR occurs at k (the number of components selected in the estimation step) as the proportion of times the bootstrap value is equal to the one obtained from the true data set.

A modification of the above method is discussed in Polymenis and Titterington (1998).

Likelihood based methodologies for model selection have also been applied to the problem of determining the number of components in a mixture. Wolfe (1971), trying to identify the number of components in a mixture of normal distributions, proposed the use of sequential tests using the LRT as a test statistic with an asymptotic $\div^2$ distribution with 4 degrees of freedom. Izenman and Sommer (1988) used this procedure in a philatelic application, justifying the use of the approximate result by the nature of their data set. They judged that, because of the multimodality of

their data, the components are well separated and thus the procedure was plausible. This was criticised by Basford *et al.* (1997). Milligan and Cooper (1985) examined the performance of several criteria for selecting the number of components in cluster analysis. They also examined the method proposed by Wolfe (1971) and suggested that this procedure tends to overestimate the number of components for medium sized samples. Soromenho (1994) compared several methods for selecting the number of components in the case of univariate normal mixtures with no more than 2 components and found that the bootstrap approach of McLachlan and the SEM method are more reliable.

Furman and Lindsay (1994a) proposed the use of moment estimates instead of ML estimates in calculating the LRT. The calculation of the ML estimates is difficult and time consuming because the EM algorithm converges slowly. Thus, they proposed the use of the moment estimates replacing the ML estimates in calculating the LRT statistic. They also showed that since the moment estimators are consistent when the sample size is large enough the difference will be negligible, and the gain in speed will be large enough. The speed is important in order to be able to obtain the distribution of the test statistic via simulation. They applied their method to normal mixtures

Maine *et al.* (1991) proposed a test based on the behaviour of the sample order statistics near the center of the distribution. The key idea is that, depending on the number of components, the differences between successive order statistics will be larger than those predicted by a simple Normal distribution and, hence, the number of components can be detected. Such a procedure, however, works only when the components are well separated.

Bayesian methodologies have been described by Aitkin *et al.* (1996) using the posterior Bayes factors, and by Richardson and Green (1997) using the newly developed reversible jump Markov Chain Monte Carlo method.

For normal models there is a large number of graphical techniques to determine the number of components. Several of them can be found in Titterington *et al.* (1985) and the references therein.

7.3 A New Method Based on the Likelihood Ratio Test

From the preceding remarks, the need for an alternative approach becomes evident. The test procedure proposed in this section aims at fulfilling this need. As will be seen, the procedure makes a sequential use of the LRT, but the null distribution of the test statistic is constructed via simulation at every stage. This is necessary since the null distribution is very sensitive to the hypotheses employed, the sample size and the closeness of the components.

The proposed approach has a dual scope: it serves as a testing hypothesis procedure for mixtures with more than one component but mainly it serves as a method of determining the number of components in a mixture. The utility of this possibility is obvious, as it gives us an insight into the structure of the population under investigation.

Consider the hypothesis

H_0 : the number of components in a Poisson mixture is k

against the hypothesis

H_1 : the number of components in the mixture is k+1.

The proposed procedure tests H_0 against H_1 sequentially for k = 1, 2, ... using the LRT statistic until H_0 is accepted for the first time at the chosen significance level. The value k_{max} of k which does not lead to the rejection of H_0 represents the optimal number of components in the Poisson mixture.

Due to the fact that the standard asymptotic result is not applicable, we adopt a resampling approach for the construction of the null distribution of the LRT statistic. The steps for carrying out the proposed test, for the case of finite Poisson mixtures, can be described as:

Set k=1

Step 1: Find the ML estimates of the parameters of the finite Poisson mixture for k and k+1 components, say $\hat{\boldsymbol{\theta}}_k$ and $\hat{\boldsymbol{\theta}}_{k+1}$ respectively and calculate the LRT statistic, say L_{obs}. Note that for k=1 the ML estimate is the sample mean.

Step 2 : Simulate B bootstrap samples of size n, (n is the sample size from the data set) from the k-finite Poisson mixture with parameter vector $\hat{\boldsymbol{\theta}}_k$, and for each bootstrap sample calculate the value of the LRT , say L_j , $j=1, \ldots, B$.

Step 3 : Estimate the á-percentile of the distribution of the test statistic by the (100a)-th order statistic from the bootstrap values L_j , $j=1, \ldots, B$. Let this percentile be C_a.

Step 4 : **If** $L_{obs} > C_a$ **then** set k=k+1 and go to step 1,

else conclude that the optimal number of components is k and stop.

As mentioned before, the procedure terminates when H_0 cannot be rejected for the first time, i.e. when there is no sufficient evidence that adding one more component will significantly improve the likelihood. We start our search with k=1, because this reduces dramatically the computational effort. Note that a similar sequential testing procedure has been considered by Aitkin *et al.* (1981), in the context of latent models. However, they only used it for a 2-class model against a 3-class one, using very few replications and without any further investigation for the performance of this approach. The work of this section is the first attempt to examine more thoroughly the procedure, and it constitutes the first application to finite mixture models.

This procedure achieves a dual goal: it reveals the number of components in the assumed mixture model, while at the same time it provides the appropriate goodness-of-fit test.

We would like to mention the attractiveness of bootstrap tests in cases where the distribution of the test statistic is not known and only asymptotic results exist. Beran (1988) has shown that bootstrap tests cannot be inferior to tests based on asymptotic results. Also, the abundance of computer resources makes bootstrap tests very practicable.

Another critical point is the inconsistency of the ML estimates when the model contains redundant components. With the proposed method such situations can be avoided since one starts with one component and increases the number of components by adding one at each step. So, the procedure will never add redundant components. In particular, it will never try to determine the ML estimates for a model with redundant components. Conditions (6.6) determine whether a new component can be added and prevent the procedure from determining the ML estimates for an inconsistent model. On the other hand, the maximised loglikelihood is consistent even when the estimates

are not, since even if the ML estimate is not unique, the maximised value of the loglikelihood is unique.

In the next section the power of the proposed method is also calculated via simulation, revealing interesting properties of the procedure. The results of the extensive simulation may be used to verify the weaknesses of the standard asymptotic result which seems inappropriate for such an approach. A point that is worth mentioning is that the null distribution of the test statistic appears to be highly dependent on the number of k under H_0 , the sample size and the data themselves, making the tabulation of such test statistic distributions impossible.

Bohning *et al.* (1994), in the concluding remarks of their paper, proposed that it might be of special interest to use such sequential testing for obtaining the 'best' number of components. Their method was of a backward search type, in the sense that they proposed to start from the model with the largest number of components and use backward elimination to find the optimal number. They also proposed to start with the semiparametric ML estimate of the mixing distribution for a finite mixture of Poisson distributions. Our method is based on a forward elimination technique which is preferable to the backward elimination technique proposed by Bohning *et al.* The reason is that the computational effort is decreased for small values of k. So, starting from k=1 the total computational effort is expected to be much less.

7.4 The Performance of the New Method

In this section an extensive simulation study of the newly proposed method is made. Two issues are of special interest:

- the first is the ability of the procedure to determine the correct number of components and
- the second is the power of the sequential tests used for obtaining the optimal number of components.

In order to use the LRT, we need the ML estimates of the mixing distribution under each of the two hypotheses. We applied the improved EM algorithm of section 3.5, using two different choices of starting values: the true values, whenever they were available, and the method which separates the interval from 0 to the maximum observed value in k equiprobable intervals. Thus, for the case with k components we set $\lambda_j^{(0)} = j\dfrac{\max_i(x_i)}{k+1}$ and $p_j^{(0)} = \dfrac{1}{k}$ for $j=1,...,k$. The convergence criterion was based on the relative increase of the likelihood between two iterations. This choice was made because we were interested in the value of the loglikelihood only.

We also employed the conditions given in (6.6) for checking if the LRT statistic is equal to 0. These conditions are very useful for applying our method since they can considerably save the computational effort needed.

In the sequel, some k-finite Poisson mixtures are considered for selected values of k (k=2,3, 4) so as to allow representation of models with well separated components, models with components close together and models that result in skew distributions. For each distribution, three sample sizes were used (n=50, 100, 500). 500 samples were subsequently generated from each distribution, for each sample size. The sequential method proposed was then applied using 500 bootstrap samples (B=500) for constructing the null distribution of each test statistic. Table 7.1 presents the relative frequencies of the numbers of components which the new method detected.

Table 7. 1

The relative frequencies of the estimated number of components among 500 simulated samples from k-finite Poisson mixtures (á=5%)

sample size	n=50					n=100					n=500				
						k=2									
	estimated number of components														
parameter vector è$_2$	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
(0.5, 1, 9)	-	0.95	0.05	-	-	-	0.95	0.05	-	-	-	0.96	0.04	-	-
(0.8, 1, 9)	-	0.92	0.08	-	-	-	0.95	0.05	-	-	-	0.96	0.04	-	-
(0.5, 1, 1.1)	0.96	0.04	-	-	-	0.93	0.07	-	-	-	0.94	0.05	0.01	-	-
(0.95, 1, 10)	0.11	0.83	0.06	-	-	-	0.93	0.07	-	-	-	0.95	0.05	-	-
						k=3									
	estimated number of components														
parameter vector è$_3$	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
(0.45,0.45,1,5,10)	-	0.62	0.36	0.01	-	-	0.39	0.58	0.02	-	-	-	0.94	0.06	-
(0.4,0.4,1,3,3.1)	0.42	0.56	0.01	-	-	0.14	0.82	0.03	-	-	-	0.96	0.04	-	-
(0.33,0.33,1,5,10)	-	0.54	0.44	0.01	-	-	0.30	0.66	0.03	-	-	-	0.94	0.06	-
						k=4									
	estimated number of components														
parameter vector è$_4$	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
(0.3,0.4,0.25,1,5,9,15)	-	0.31	0.61	0.08	-	-	0.09	0.78	0.13	-	-	-	0.59	0.38	0.03
(0.3,0.3,0.2,1,1.2,5,9)	-	0.78	0.21	0.01	-	-	0.68	0.31	0.01	-	-	0.17	0.78	0.03	0.02
(0.25,0.25,0.25,1,5,10,15)	-	0.17	0.76	0.07	-	-	0.02	0.86	0.12	-	-	-	0.59	0.40	0.01

Table 7.1 reveals that the method is quite successful in determining the number of components when these are not close and the sample size is large enough. How large the sample size must be depends on the value of k. So, for k=3 a sample size of 500 is sufficient for the accurate determination of the number of components. For k=4 larger sample sizes are needed. Clearly, when the components are very close, the method cannot distinguish between them. On the other hand, components with small mixing probabilities are usually ignored, especially in the case of small sample sizes. It is interesting that the method seldom overestimates the number of components. For small sample sizes it performs better for models with small numbers of components. This may be connected with the high variances of the ML estimates for finite Poisson mixtures with not well separated components and small sample size, first reported by Hasselblad (1969). For our simulation purposes, only cases plausible in practice and small sample sizes were considered. Selecting cases with extraordinarily large separation between the components would only lead to more impressive results but of little practical interest, as most often count data consist of small positive integers.

The sequential nature of the tests employed makes the calculation of the power of the method (in the usual sense used in hypothesis testing) very difficult. The simulation results reported in Tables 7.1a-c, however, can also be regarded as revealing the power of the proposed method.

The power of each separate test proposed is also examined. Thode *et al.* (1988) and Mendell *et al.* (1991, 1993), have examined the power of the LRT for testing one component versus two components in normal mixtures with equal variances via simulation. Recently, Berdai and Garrel (1996) examined the power of the LRT deriving an asymptotic distribution. All the authors agree in that the power of the test is susceptible to the sample size and to the closeness of the components .

In order to investigate the power of the proposed method, the empirical power (see definition 6.1) of the test for k components versus k+1 components was examined for k = 1, 2, 3. Exact calculations require knowledge of the distribution of the test statistic under the alternative distribution which is not easily obtainable. The level of significance was á=5%. The critical value of each test was calculated via simulation of 2000 samples of given size from the null distribution. For each value

of k, several distributions were chosen so as to represent various cases. For every distribution examined, 4 alternatives were considered. The alternative distributions were chosen so as to have the same mean with the distribution under the null hypothesis. The reason is that when applying the test to real data sets the ML estimates for k-finite mixtures must satisfy the first moment equation whatever the value of k. This is also true for normal mixtures which makes values of the reported power in Mendell *et al.* (1991) irrelevant.

For the case where k=1, the null distributions used were Poisson distributions with parameters ë= 1, 3, 5, 10, respectively. Ôhen, the vectors of parameters for the 2-finite mixture alternatives were:

(1A) (0.5, 0.95ë, 1.05ë),

(1B) (0.5, 0.5ë, 1.5ë),

(1C) (0.8, 0.8ë, 1.8ë) and

(1D) (0.2, 0.8ë, 1.05ë).

Alternative (1A) is very close to the null distribution while (1C) and (1D) result in distributions more skew to the left and to the right, respectively.

For k=2, the distributions considered in the null hypothesis had vectors of parameters of the form $(p_1, ë_1, ë_2)$:

(2a) (0.5, 1, 5),

(2b) (0.8, 3, 11),

(2c) (0.5, 2, 2.2),

(2d) (0.5, 5, 15).

The four alternatives for each null hypothesis considered were of the form:

(2A) $(0.5p_1, 0.5p_1, 0.95ë_1, 1.05ë_1, ë_2)$,

(2B) $(0.5p_1, 0.5p_1, 0.5ë_1, 1.5ë_1, ë_2)$,

(2C) $(ðp_1, ð, 0.95ë_1, (ë_1+ë_2)/2, ë_2)$,

(2D) ($\eth p_1$, $\eth p_2$, $0.95ë_1$,$ë_2$,$1.5ë_2$) and

(**2E)** (0.33,0.33, 1, 1+a,$ë_2$+1) ,

where ð is the probability assigned to the third component so that the mean does not change and a is chosen so that the alternative distribution can have the same mean as the null distribution. Again (2A) differs very little from the null distribution, (2B) differs more, while (2C) and (2D) add the new component at the left and the right tail respectively.

Similarly, for k=3 the distributions used under the null hypothesis had vectors of parameters of the form (p_1, p_2, $ë_1$, $ë_2$, $ë_3$):

(3a) (0.33, 0.33, 1, 5, 12) ,

(3b) (0.8, 0.1, 1, 5, 12),

(3c) (0.1 0.4, 1, 5, 12),

(3d) (0.5, 0.25, 1, 8, 8.5) and

(**3e**) (0.33,0.33,1,10,20).

The alternatives considered for each of them were of the form:

(3A) ($0.5p_1$, 0.5 p_1, p_2, $0.95ë_1$, 1.05 $ë_1$, $ë_2$, $ë_3$),

(3Â) ($0.5p_1$, 0.5 p_1, p_2, $0.5ë_1$, 1.5 $ë_1$, $ë_2$, $ë_3$),

(3C) (p_1, p_2, $0.5p_3$, $ë_1$, $ë_2$, $0.5ë_3$, 1.5 $ë_3$),

(3D) ($\eth p_1$, $\eth p_2$, $\eth p_3$, $ë_1$, $ë_2$, $ë_3$, 1.5 $ë_3$) and

(3E) (0.25,0.25,0.25,1,1+a,1+2a,$ë_3$+1),

where a and ð are defined as previously. Again, (3A) differs very little from the null distribution, (3B) differs more, while (3C) and (3D) add the new component between the 2nd and the 3rd component and at the right tail respectively.

Tables 7.2a-c contain the empirical power for all the cases.

Table 7.2a
The empirical power of the LRT for testing k=1 versus k=2 (á=5%).

Null distribution	alternative	sample size n=50	n=100	n=500	n=1000	n=2000
ë=1	1A	0.063	0.068	0.064	0.050	0.048
	1B	0.340	0.526	0.974	0.999	1.000
	1C	0.225	0.327	0.753	0.936	0.997
	1D	0.081	0.089	0.087	0.080	0.095
ë=3	1A	0.041	0.048	0.039	0.031	0.035
	1B	0.868	0.989	1.000	1.000	1.000
	1C	0.538	0.793	1.000	1.000	1.000
	1D	0.061	0.070	0.094	0.100	0.147
ë=5	1A	0.033	0.039	0.035	0.027	0.034
	1B	0.992	1.000	1.000	1.000	1.000
	1C	0.802	0.972	1.000	1.000	1.000
	1D	0.062	0.074	0.132	0.171	0.294
ë=10	1A	0.027	0.034	0.042	0.035	0.060
	1B	1.000	1.000	1.000	1.000	1.000
	1C	0.988	1.000	1.000	1.000	1.000
	1D	0.073	0.107	0.320	0.511	0.813

Table 7.2b
The empirical power of the LRT for testing k=2 versus k=3 (á=5%).

Null distribution	alternative	sample size n=50	n=100	n=500	n=1000	n=2000
2a	2A	0.046	0.059	0.065	0.052	0.044
	2B	0.052	0.090	0.222	0.314	0.514
	2C	0.027	0.035	0.026	0.006	0.000
	2D	0.066	0.080	0.078	0.063	0.049
	2E	0.104	0.197	0.544	0.764	0.949
2b	2A	0.049	0.067	0.051	0.039	0.040
	2B	0.393	0.697	0.999	1.000	1.000
	2C	0.060	0.074	0.084	0.080	0.075
	2D	0.108	0.123	0.166	0.174	0.190
	2E	0.152	0.192	0.347	0.502	0.755
2c	2A	0.034	0.038	0.077	0.085	0.077
	2B	0.097	0.119	0.190	0.287	0.460
	2C	0.002	0.001	0.000	0.000	0.000
	2D	0.011	0.011	0.003	0.002	0.002
	2E	0.012	0.020	0.045	0.033	0.018
2d	2A	0.051	0.065	0.050	0.040	0.037
	2B	0.422	0.723	1.000	1.000	1.000
	2C	0.048	0.056	0.061	0.044	0.044
	2D	0.084	0.100	0.104	0.113	0.119
	2E	0.105	0.130	0.198	0.272	0.447

Table 7.2c
The empirical power of the LRT for testing k=3 versus k=4 (á=5%).

Null distribution	alternative	sample size n=50	n=100	n=500	n=1000	n=2000
3a	3A	0.032	0.068	0.097	0.085	0.047
	3B	0.025	0.064	0.183	0.232	0.355
	3C	0.015	0.030	0.058	0.053	0.032
	3D	0.092	0.180	0.551	0.764	0.941
	3E	0.029	0.068	0.225	0.302	0.462
3b	3A	0.050	0.067	0.112	0.121	0.109
	3B	0.063	0.123	0.424	0.632	0.831
	3C	0.013	0.027	0.104	0.126	0.126
	3D	0.050	0.123	0.528	0.766	0.937
	3E	-	-	-	-	-
3c	3A	0.021	0.050	0.107	0.082	0.036
	3B	0.017	0.040	0.114	0.138	0.142
	3C	0.016	0.031	0.008	0.001	0.000
	3D	0.032	0.066	0.092	0.087	0.115
	3E	0.025	0.043	0.081	0.070	0.063
3d	3A	0.049	0.060	0.071	0.063	0.074
	3B	0.080	0.125	0.368	0.575	0.808
	3C	0.001	0.003	0.001	0.000	0.000
	3D	0.008	0.017	0.064	0.140	0.303
	3E	0.041	0.058	0.113	0.122	0.129
3e	3A	0.046	0.057	0.064	0.047	0.041
	3B	0.082	0.152	0.431	0.644	0.893
	3C	0.036	0.045	0.051	0.036	0.013
	3D	0.146	0.252	0.297	0.537	0.761
	3E	0.244	0.471	0.962	0.999	1.000

For testing a one component mixture versus a two component mixture, the power of the test increases with the distance of the components. This result is similar to the one obtained for normal mixtures by Mendell *et al.* (1991) and it was not unexpected. For testing k=2 versus k=3, the power is increased only when the sample size is large and the components are well separated. Adding a well separated new component, but with a small probability, does not improve the power of the test. This

is the case for k=3 versus k=4, too. Concluding, we can say that the LRT applied to the general case k=m versus k=m+1 has a low power when the components are not well separated and when one of the components has a small mixing probability. As the value of m increases, the sample size required for obtaining a specific power increases very much. This result verifies the behaviour of the method for the simulated cases of Tables 7.1a-c. As far as the asymptotic distribution of the test statistic is concerned, the $\div^2$ form does not seem plausible. On the other hand, the null distribution depends highly on the value of k and the sample size used.

7.5 Applications

In the present section the proposed procedure for determining the number of components in the case of a finite Poisson mixture is illustrated by two examples referring to real datasets.

Example 7.1 The first example concerns the data considered by Bohning *et al.* (1992), among others. The data refer to the number of death notices for women aged 80 and over, in the Times newspaper for each day in the 3-year period from 1910 to 1912. The frequencies can be found in Table 7.3.

The mean of this data set is equal to 2.156 while the variance is equal to 2.607. So, overdispersion is present, and a finite Poisson mixture model is a plausible assumption.

Indeed, the simple Poisson case provides a very poor fit to the data as shown in Table 7.3. Therefore, fitting the data by a Poisson mixture might be more appropriate. The results of Table 7.3 come in support of this assumption.

We used the new procedure of section 7.3 in order to determine the number of components. The method provides evidence for a 2-finite Poisson mixture as the underlying distribution, which indeed fits the data very well as judged by the value of the $\div^2$ test statistic and the entries of Table 7.4.

Table 7.3
Observed and fitted frequencies for the number of death notices for women aged 80 and over in Britain for the period 1910-1912. The asterisks indicate the grouping adopted for calculating the $\div^2$ values.

x	observed	Poisson	2-finite Poisson mixture	3-finite Poisson mixture
0	162.00	126.79	160.92	161.23
1	267.00	273.47	271.35	271.41
2	271.00	294.92	262.30	262.08
3	185.00	212.04	191.25	191.05
4	111.00	114.34	114.21	114.16
5	61.00	49.32	57.52	57.55
6	27.00	17.73	24.83	24.87
7	8.00	5.46	9.32	9.35
8	3.00	* 1.47	* 3.08	* 3.09
9	1.00	* 0.35	* 0.91	* 0.91
$\div^2$ **value**		24.96	1.494	1.492
df		7	5	3

Table 7.4
Sequential testing results for the data in Table 7.3

k	LRT statistic	p-value
1	22.904	0
2	0.038	0.339

Column 1 of Table 7.4 contains the values of k, the number of the components in the mixture. Column 2 contains the value of the maximised loglikelihood for a model with k components, while the values of the test statistic for testing m=k against m=k+1 can be seen in the third column. Last column contains the associated p-values calculated via simulation. Using the bootstrap approach described previously, we constructed the null distribution of the test statistic for various values of k, using 10000 bootstrap samples. Based on Table 7.4 we reject the hypothesis that the simple Poisson distribution (k=1) fits the data, but we cannot reject the hypothesis that a 2-finite Poisson distribution fits the data. So, we may regard that k=2, i.e. that the number of components is two.

Looking at figure 7.1 it is worth noting that the distribution of the LRT depends on the value of k as specified by the null hypothesis. This is in contradiction with the standard technique for the LRT which assumes that, asymptotically, the distribution of the test statistic is the same at every step. Note also that the distribution of the test statistic tends to be concentrated towards smaller values as k increases. Finally, it can clearly be seen that the $\div^2$ approximation is very poor.

We may deduce, therefore, that the use of the $\div^2$ can lead to invalid conclusions, and hence should be avoided. Bohning *et al.* (1994) have come to the same conclusion for a variety of models in the case k=1.

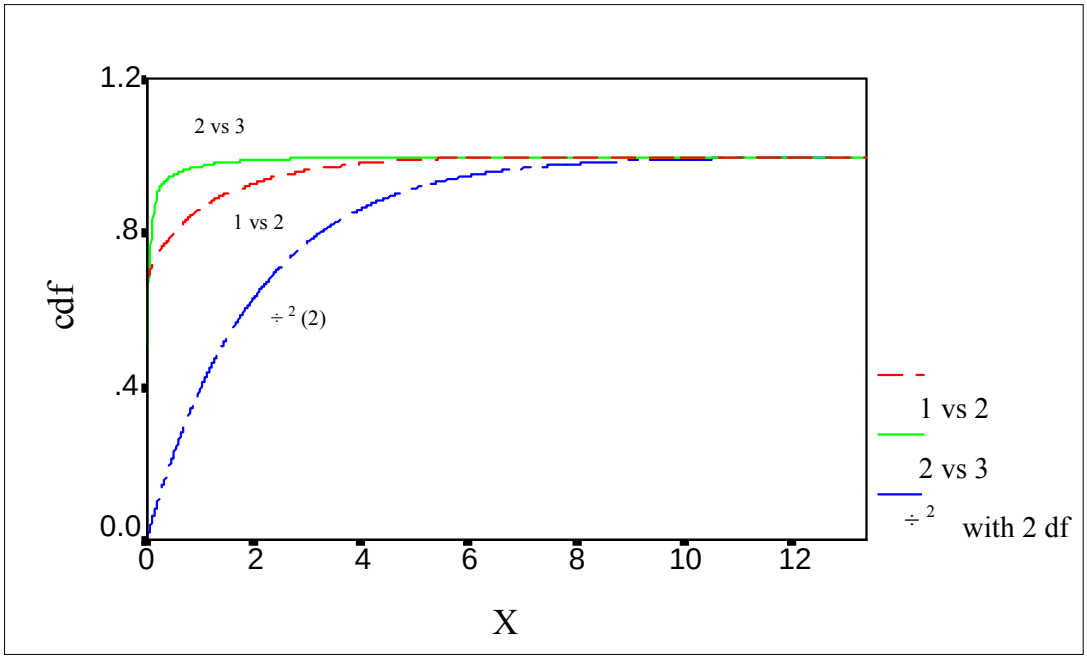


Figure 7.1 The cumulative distribution function for the test statistic for testing H_0 : k=1 vs H_1 : k=2 and H_0 : k=2 vs H_1 : k=3 for the data of the first example and that of a χ^2 distribution with 2 degrees of freedom. Clearly the form of the distribution depends on the value of k.

Some descriptive statistics from the simulations can be seen in Table 7.5. Again it is clear that the $\div^2$ approximation is invalid. At the bottom of Table 7.5, we can see the conditional descriptive statistics, conditional on non zero values of the test statistic. *P(0)* represents the proportion of zero values.

Table 7.5
Simulation results for the data of Table 7.3

	k=1 vs k=2	k=2 vs k=3	$\div^2(2)$
80th percentile	0.7313	0.1054	3.2189
90th percentile	1.8237	0.5883	4.6053
95th percentile	3.0141	1.6851	5.9919
97.5th percentile	4.6726	3.0575	7.3778
99th percentile	7.0746	4.3817	9.2103
P(0)	0.561	0.490	
mean	0.5679	0.2598	2
median	0	0.0001	
standard deviation	1.3765	0.8714	2
conditional mean	1.2936	0.4853	
conditional median	0.6001	0.0419	
conditional standard deviation	1.8380	1.1443	

Example 7.2 Consider now the data given in Table 7.6 on the number of accidents incurred by 414 machinists over a period of three months, taken from the classical paper of Greenwood and Yule (1920), and analysed by several authors. Note that this data set is more skew than the previous. Again, the fit provided by the simple Poisson distribution is very poor as seen in Table 7.6. This data set shows a large overdispersion. The mean equals 0.483, while the variance is 1.010, i.e. more than twice the mean. A notable improvement is achieved by fitting mixtures of Poisson models.

Table 7.6
Observed and fitted frequencies for the number of accidents over a period of three months for 414 machinists. The asterisks indicate the grouping adopted for calculating the $\div^2$ values.

x	observed	Poisson	2-finite Poisson mixture	3-finite Poisson mixture	4-finite Poisson mixture
0	296.00	255.41	294.10	296.10	297.34
1	74.00	123.36	78.46	74.40	75.10
2	26.00	29.76	20.36	24.79	24.85
3	8.00	* 4.80	10.52	8.99	8.24
4	4.00	* 0.58	5.99	4.40	3.88
5	4.00	* 0.06	2.86	2.60	2.27
6	1.00	* 0.00	* 1.15	* 1.47	* 1.27
7	0.00	* 0.00	* 0.39	* 0.73	* 0.62
8	1.00	* 0.00	* 0.12	* 0.32	* 0.27
$\div^2$		57.812	4.7045	3.2062	3.1264
df		2	3	1	-

Again in Table 7.7 we summarise the results of our approach. The 3-component model is an indisputable choice. Note, however that, had we erroneously used the $\div^2$ approximation, we would have chosen the 2-component model because the observed value 3.122 of the LRT statistic would not have led to the rejection of the null hypothesis if the critical value of the $\div^2$ distribution has been used.

Table 7.7
Sequential testing results for the data in Table 7.6

k	LRT statistic	p-value
1	88.068	0
2	3.122	0.033
3	0.094	0.216
4	0.024	0.185

Table 7.8 contains again some descriptive measures derived from the simulations. Again, the inadequacy of the $\div^2$ approximation is clear.

Table 7.8
Simulation results for the data in Table 7.6

	k=1 vs k=2	k=2 vs k=3	k=3 vs k=4	$\div^2(2)$
80th percentile	0.7585	0.5512	0.0513	3.2189
90th percentile	1.9376	1.6029	0.1760	4.6053
95th percentile	3.4370	2.6861	0.4899	5.9919
97.5th percentile	4.5530	3.7990	0.9936	7.3778
99th percentile	6.2881	5.2353	1.8350	9.2103
P(0)	0.6540	0.5960	0.5380	
mean	0.5740	0.4497	0.0945	2
median	0	0	0	
stdev	1.3155	1.0971	0.3748	2
conditional mean	1.3251	1.1132	0.1822	
conditional median	0.6492	0.5383	0.0241	
conditional standard deviation	1.7321	1.4971	0.5048	

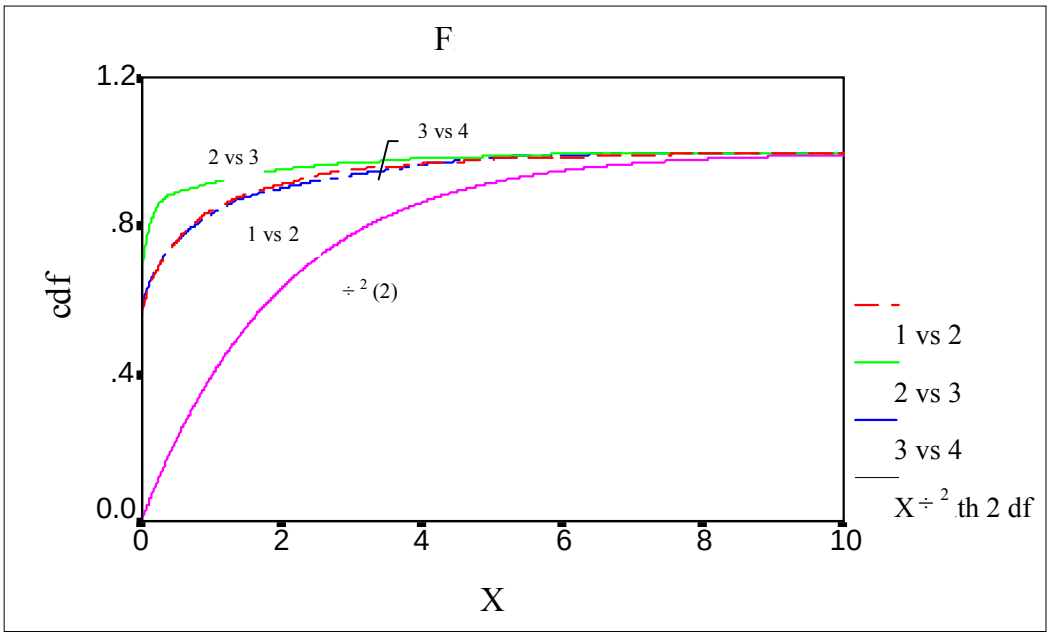


Figure 7.2 The cumulative distribution function for the test statistic for testing H_0 : k=1 vs H_1 : k=2, H_0 : k=2 vs H_1 : k=3 and H_0 : k=3 vs H_1 : k=4 for the data of the second example and that of a χ^2 distribution with 2 degrees of freedom. Again, the form of the distribution clearly depends on the value of k.

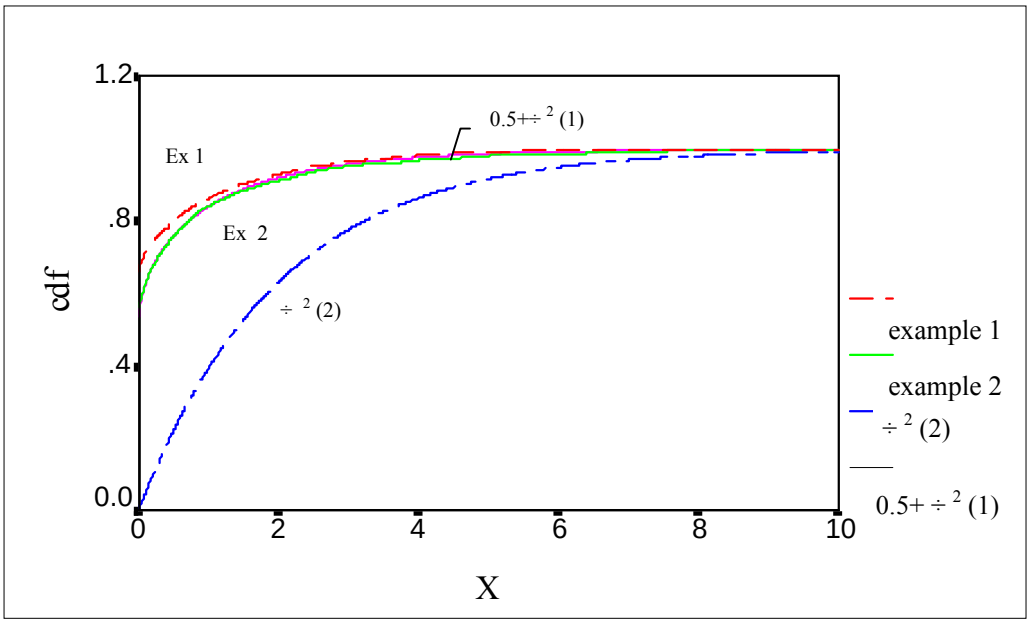


Figure 7.3 The cumulative distribution function for the test statistic for testing H_0 : k=1 vs H_1 : k=2 for the two examples, and those of r a $\div^2$ with 2 df and a mixture of a degenerate distribution at 0 and a $\div^2$ with 1 df. Clearly, the latter is very close to the simulated distribution, especially towards the tail of the distribution.

In Figure 7.2 we can see the distribution of the test statistic for the hypotheses tested. Clearly, the distributions differ markedly from the $\div^2$ distribution with 2 degrees of freedom and there is a difference between the distributions corresponding to different values of k in the null hypothesis.

Figure 7.3 shows the cumulative distribution function of the test statistic for testing H_0: k=1 vs H_1 : k=2 for the two datasets. The distributions are very similar, verifying the results reported in the previous chapter about the distribution of the LRT.

In order to assess the performance of the newly proposed method, we calculated the empirical power of the test procedures involved. As mentioned before, this is defined as the proportion of times we rejected the null hypothesis when the data actually came from the alternative distribution. So, for both the examples, we used as critical values for given á the corresponding á-percentiles of the null distribution constructed via simulation. 10000 samples were generated from the distribution in H_1, namely the distribution with parameters the ML estimates for a model with the number of components specified in H_1. For each sample, the LRT statistic was calculated so as to construct the distribution of the test statistic under the alternative hypothesis. The proportion of times H_0 was rejected for a given level of significance á are reported in Table 7.9. As can be easily seen, the LRT performs well only for the case k=1 vs k=2. In the first example the test lacks power for testing 2 components versus 3 components since the components are not well separated. This holds true for the second example as well. For both examples, the tests have low power when testing for points which are redundant. Generally speaking, it can be noted that the test has a lower performance when it is used to detect components that are very close. This is usually true for models with a large number of components as the new added point is usually very close to the previously estimated points. Note that the null distribution of all the test statistics is highly skewed to the right.

Table 7.9
Asymptotic Power calculation (á denotes the significance level).

á=	0.10	0.05	0.025	0.01
Example 1				
k=1 vs k=2	1.000	0.999	0.996	0.991
k=2 vs k=3	0.117	0.054	0.027	0.015
Example 2				
k=1 vs k=2	1.000	1.000	1.000	1.000
k=2 vs k=3	0.560	0.430	0.318	0.207
k=3 vs k=4	0.055	0.023	0.008	0.003

Based on the above results on both real and simulated data, the method presented in this chapter does not seem to overestimate the number of components in the mixture. This is the consequence of the fact that in a model with too many components, two or more components are essentially duplicates, and thus the improvement of the loglikelihood is negligible.

7.6 Conclusions

The performance of the newly proposed method seems to be quite satisfactory, as the simulations revealed. This method is of much general interest and it can be used for other finite mixtures, like normal or exponential mixtures. The key-idea is to use sequentially the LRT, via bootstrap simulations. These extensive simulations are computationally intensive, but the impact of powerful computer resources makes feasible the applicability of such methods. The conditions derived in chapter 6, can save a lot of computational effort.

The Hellinger deviance test can also be considered as a counterpart of the sequential LRT. Likelihood based methods, usually treat the extreme observations (outliers) as coming from a further component. As seen in chapters 5 and 6, an observation not so far from the main body of the dataset can have a large influence on both the ML estimates and the LRT and, hence, this increase may be significant. Till now we have not considered the HDT as a method for determining the number of components but the intrinsic robustness of Hellinger based methods is a promising characteristic. The derivation of a sequential HDT is still an open problem. We hope to be able to report results soon.

The methodology proposed in this chapter has been applied to finite Poisson mixtures. The extension of the reported results to other kind of finite mixtures is obvious. We may apply the sequential LRT for finite normal mixtures or latent class models, for example. This approach may also be of particular interest in the area of cluster analysis, see, e.g., McLachlan (1992, pp 22).

Finally, it should be emphasised that the method introduced in this chapter, can lead to the development of other procedures for testing hypotheses for mixtures with more than one component, which are very few in the literature.

Chapter 8

Conclusions and Open Problems

This thesis contains new results concerning Poisson mixtures finite or not. Extensive reviews of existing results are given covering the literature on such models up to now. Some important new results are also provided.

There are a lot of interesting problems which could not have been addressed in this thesis. The wide applicability of mixed Poisson models, makes it interesting to mention some of them in order to stimulate the interested reader for further research. Several cases are still part of ongoing research.

Chapter 1 presented proneness and contagion models. These two distinct models, starting from quite different assumptions about the underlying situation, lead to the same distributional form. Such an example is the negative binomial distribution which can be derived via both a proneness and a contagion model (and in fact via several other models). It remains an open problem to distinguish between these two models. In this case, as in all other cases in which other mixed Poisson distributions arise, additional information is needed. Using the exact times of the occurrence of the events does not facilitate the solution of the problem (see, e.g. Cane (1977) and Xekalaki (1983)). Looking at the data arranged in two consecutive time periods can provide additional insight (see Xekalaki (1984)). So, it remains an open problem to find a procedure for distinguishing between these two models (and perhaps between some other models).

Chapter 2 contains material about mixed Poisson distributions. In practice, only a few of these results have been used. The reason is the complicated form of their probability function, and the difficulty in deriving estimates for their parameters. Estimation procedures which can simplify the estimation of the parameters of mixed

Poisson distributions are interesting. Consider the negative binomial distribution, a prominent member of the class of mixed Poisson distributions. The ML estimators of its parameters are hard to be derived, and special numerical methods are needed. The same is true for the majority of the mixed Poisson distributions described.

It would also be interesting to examine more thoroughly some important (but yet rather ignored) members of this family. As seen in this thesis, the mixed Poisson distribution resembles the shape of the mixing distribution. So, continuous distributions which can describe particular forms of data can be used as mixing distributions, for deriving discrete analogues. Such a special distribution is the Poisson Lognormal distribution. Its probability function cannot be written in a closed form and a recurrence relation is not available. This leads to the limited applicability of this distribution which can describe data sets that exhibit high skewness. However, in recent years, the increased computational power can help overcome such problems. Further research is needed so as to examine more thoroughly the properties of a large number of mixed Poisson distributions and to make a comparative study of their performance.

Chapter 3 treats the ML estimation method for finite Poisson mixtures and the more general form of semiparametric ML estimation. Even when the mixing distribution is continuous, we can estimate it only by a finite step distribution. The EM algorithm is a promising procedure for ML estimation when the number of support points is known. One method for substantially improving the speed of the EM algorithm was proposed. It was also tried to find better initial values, as well as better stopping rules for the algorithm. All these aspects can improve the algorithm. Further improvements are needed. Several choices of procedures for finding good initial values, as well as better stopping rules for the algorithm have also been considered. As far as semiparametric ML estimation is concerned, it was shown that the proposed methods are not efficient for the case of Poisson mixtures. The results about the M2 type of samples, when additional information about the observations is available, encourages the search for this additional information, although this can be expensive. The gain is very large even with small additional information. This chapter contains, also, interesting extensions of the applications of the EM algorithm for finite mixtures to problems which can be considered as mixture problems, such as in the case of distributions with added zeroes. It would be useful to extend such applications

to other models, so as to apply the simple EM algorithm. Such problems are the general problem of weighted distributions, mixture models with non identical components, models containing restrictions such as models with added probability in some points, outlier detection problems etc.

Chapter 4 discourages the use of the moment method of estimation. The moment estimates very often do not exist and when they exist they are inferior to the ML estimates. A modification of the classical moment estimation was proposed by replacing higher order moments with other functionals. So, alternative estimation methods were considered which are very useful in particular applications. The zero frequency method was presented, which is very interesting when special attention must be paid to the zero frequency.

Chapters 5 and 6 provide robust methodologies alternative to the likelihood based methodologies, covering estimation, testing of hypothesis and diagnostic matters. These are based on the Minimum Hellinger Distance. Both these chapters contain almost entirely new results. It is very interesting that Minimum Hellinger methodologies are preferable since they combine high efficiency with strong robustness. These results could be extended to cover other problems such as mixed Poisson regression problems, estimation and hypothesis testing for mixtures from other distributions, or problems in cluster analysis.

As Lindsay (1994) showed, several other distances can be considered for minimum distance estimation. The MHD used in this thesis is known to have useful robustness properties to outliers, while other distances are more appropriate for other kinds of departure form the assumed model, such as the presence of inliers etc. Developing such methods would be very interesting. In this thesis only the case of data contamination was examined. The case of model deviation has not been dealt with despite its practical interest. For example, it would be interesting to examine the behaviour of estimation methods when the component distributions deviate from the Poisson forms considered in this thesis.

Chapter 7, treats a problem which does not seem to have a clear answer: how many components do exist in a mixture? This problem, and the similar in nature problem on the determination of the number of clusters, is in fact insolvable. A new method was developed, which utilises the LRT statistic sequentially, in order to find

the number of components for which the addition of another component will not improve significantly the likelihood. It is interesting to examine if using the test based on the Hellinger deviance would improve our insight. The sequential LRT procedure could be generalised by allowing for covariates in our model. It should be emphasised that we used this new method for Poisson mixtures. Clearly, the method is also applicable to other families of mixtures.

Only univariate Poisson mixtures were examined in this thesis. It would be interesting to examine bivariate mixed Poisson distributions, too. Bivariate extensions can provide a better insight into the situation under consideration, which makes them promising models for real applications. Unfortunately, since the bivariate Poisson distribution has 3 parameters, several distinct mixed bivariate Poisson distributions can be constructed.

Mixture models have become very fashionable nowadays. The availability of computer resources has removed any obstacles to the application of mixture models and now a variety of distinct fields of statistical methodology uses mixture models. A lot of research on mixture models is expected to be carried out in the next years and we hope that the problems covered in this thesis will help in enhancing the existing knowledge about mixtures.

APPENDIX

The Jacknife Estimator of the Variance

The jacknife is a useful technique for estimating the bias and the standard errors of an estimate. The jacknife focuses on the samples that leave out one observation at a time, i.e. we calculate the estimates for all the samples resulting when we drop an observation each time. More formally if the full sample is $\mathbf{X} = (x_1, x_2, \ldots, x_n)$, then we represent the i-th jacknife sample as: $\mathbf{X}_{(i)} = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$, i.e. the i-th jacknife sample is the sample which results form the full sample if we leave out the i-th observation.

Suppose that we have an estimator $\hat{\theta} = s(\mathbf{X})$ based on the full sample. Then we can derive the i-th jacknife estimate of θ as $\hat{\theta}_{(i)} = s(\mathbf{X}_{(i)})$, i.e. the estimate derived form the i-th jacknife sample. Then the jacknife estimate of standard error is calculated as

$$s_{jack} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \hat{\theta}_{(.)}\right)^2} \quad ,$$

where $\hat{\theta}_{(.)} = \frac{\sum_{i=1}^{n}\hat{\theta}_{(i)}}{n}$. Note that the factor (n-1) is introduced to account for the similarity of the jacknife samples.

Jacknife estimators are known to have less bias than standard estimators. In many cases exact standard errors are not easily derived and thus the jacknife standard errors can be derived more easily.

In all the applications in this thesis, the jacknife standard errors were calculated. For more details on jacknife the reader is referred to Efron and Tibshirani (1993).

References

Abramowitz, M. and Stegum, I.A. (1965). *Handbook of Mathematical Functions*. New York, Dover.

Adell, J. and de la Cal, J. (1993). On the Uniform Convergence of Normalised Poisson Mixtures to Their Mixing Distribution. *Statistics and Probability Letters,* 18, 227-232.

Agha, M. and Ibrahim, M. (1984). Algorithm AS203: Maximum Likelihood Estimation of Mixtures of Distributions. *Applied Statistics,* 33, 327-332.

Ahmad, M., Giri, N. and Sinha, B.K. (1983). Estimation of the Mixing Proportion of Two Distributions. *Sankhya,* A 45, 357-371.

Aitkin, M. and Aitkin, I. (1996). An Hybrid EM/Gauss-Newton Algorithm for Maximum Likelihood in Mixture Distributions. *Statistics and Computing,* 6, 127-130.

Aitkin, M., Anderson, D. and Hinde, J. (1981). Statistical Modelling of Data on Teaching Styles. *Journal of the Royal Statistical Society,* A 144, 419-461.

Aitkin, M., Finch, S., Mendell, N. and Thode, H. (1996). A New Test for the Presence of a Normal Mixture Distribution Based On the Posterior Bayes Factor. *Statistics and Computing,* 6, 121-125.

Aitkin, M. and Rubin, D. (1985). Estimation and Hypothesis Testing in Finite Mixture Models. *Journal of the Royal Statistical Society,* B 47, 67-75.

Aitkin, M. and Wilson, T. (1980). Mixture Models, Outliers and the EM Algorithm. *Technometrics,* 22, 325-331.

Albrecht, P. (1980). On the Correct Use of the Chi-Square Goodness of Fit Test. *Scandinavian Actuarial Journal,* 7, 149-160.

Albrecht, P. (1982). On Some Statistical Methods Connected with the Mixed Poisson Process. *Scandinavian Actuarial Journal,* 9, 1-14.

Albrecht, P. (1984). Laplace Transforms, Mellin Transforms and Mixed Poisson Processes. *Scandinavian Actuarial Journal,* 11, 58-64.

Al-Hussaini, E.K. and Abd-EL-Hakim, N.S. (1990). Estimation of Parameters of the Inverse Gaussian - Weibull Mixture Model. *Communications in Statistics-Theory and Methods*, 19, 1607-1622.

Al-Hussaini, E.K. and El-Dab, A.K. (1981). On the Identifiability of Finite Mixtures of Distributions. *IEEE Transactions on Information Theory*, 27, 664-668.

Anderson, J. and Siddiqui, M. (1994). The Sampling Distribution of the Index of Dispersion. *Communication in Statistics-Theory and Methods,* 23, 897-911.

Anscombe, F.J. (1950). Sampling Theory of the Negative Binomial and Logarithmic Series Distributions. *Biometrika,* 37, 358-382.

Aragon, J., Eberly, D. and Eberly, S. (1992). Existence and Uniqueness of the Maximum Likelihood Estimator for the Two-parameter Negative Binomial Distribution. *Statistics and Probability Letters,* 15 , 375-379.

Arbous, A.G. and Kerrich, J.E. (1951). Accident Statistics and the Concept of Accident-Proneness. *Biometrics,* 7, 340-432.

Arbous, A.G. and Kerrich, J.E. (1954). New Techniques for the Analysis of Absenteism Data. *Biometrika,* 41, 77- 90.

Atkinson, A.C. and Yeh, L. (1982). Inference for Sichel's Compound Poisson Distribution. *Journal of the American Statistical Association,* 77, 152-158.

Atwood, L.D., Wilson, A.F, Elston, R.C. and Bailey-Wilson, J.E. (1992). Computational Aspects of Fitting a Mixture of Two Normal Distributions Using Maximum Likelihood. *Communications in Statistics- Simulation and Computation,* 21, 769-781.

Atwood, L.D., Wilson, A.F., Bailey-Wilson, J.E., Carruth, J.N. and Elston, R.C. (1996). On The distribution of the Likelihood Ratio Tests Statistic for a Mixture of two Normal Distributions. *Communications in Statistics- Simulation and Computation,* 25, 733-740.

Barndorff-Nielsen, O.E. (1965). Identifiability of Mixtures of Exponential Families. *Journal of Mathematical Analysis and Applications,* 12, 115-121.

Barndorff-Nielsen, O.E., Kent, J. and Sorensen, M. (1983). Normal Variance-Mean Mixtures and z-Distributions. *International Statistical Review*, 50, 145-159.

Barndorff-Nielsen, O.E., Blaesild, P. and Seshardi, V. (1992). Multivariate Distributions with Generalised Inverse Gaussian Marginals and Associated Poisson Mixtures. *Canadian Journal of Statistics,* 20, 109-120.

Bartlett, M. and Mc Donald, P. (1968). Least Squares Estimation of Distribution Mixtures. *Nature E.S.I. London,* 217, 195-196.

Basford, K.E. and McLachlan, G. (1985). Likelihood Estimation with Normal Mixtures Models. *Applied Statistics,* 34, 282-289.

Basford, K.E., Greenway, D.R., McLachlan, G.J. and Peel, D. (1997). Standard Errors of Fitted Component Means of Normal Mixtures. *Computational Statistics,* 12, 1-18.

Basford, K.E., McLachlan, G.J. and York, M.G. (1997). Modelling the Distribution of Stamp Paper Thickness via Finite Normal Mixtures: The Hidalgo Stamp Issue of Mexico Revistited. *Journal of Applied Statistics,* 27, 169-181.

Basu, A., Basu, S. and Chaudhuri, G. (1997). Robust Minimum Divergent Procedures for Count Data. *Sankhya,* B 59, 11-27.

Basu, A. and Lindsay, B. (1992). The Iteratively Reweighted Estimating Equation in Minimum Distance Problems. *Technical Report 92-02, Center for Likelihood Studies, Department of Statistics, Pensylvania State University* .

Basu, A. and Lindsay, B. (1994). Minimum Disparity Estimation for Continuous Models: Efficiency, Distributions and Robustness. *Annals of the Institute of Statistical Mathematics*, 46, 683-705.

Basu, A. and Sarkar, S. (1994). The Trade-off Between Robustness and Efficiency and the Effect of Model Smoothing in Minimum Disparity Inference. *Journal of Statistical Computation and Simulation,* 50, 173-185.

Bates, G.and Neyman, J. (1952a). Contributions to the Theory of Accident Proneness Part II: True Or False Contagion? *University of California Publications in Statistics, 1952*, 255-275 .

Bates, G.and Neyman, J. (1952b). Contributions to the Theory of Accident Proneness Part I: An Optimistic Model of the Correlation Between Light and Severe Accidents *University of California Publications in Statistics, 1952*, 215-253 .

Beall, G. and Rescia, R.R. (1953). A Generalization of Neyman's Contagious Distributions. *Biometrics,* 9, 354-386.

Behboodian, J. (1970). On a Mixture of Normal Distributions. *Biometrika,* 57, 215-217 .

Beran, R.J. (1977). Minimum Hellinger Distance Estimates for Parametric Models. *Annals of Statistics,* 5, 445-463.

Beran, R.J. (1988). Prepivoting Test Statistics: a Bootstrap Review of Asymptotic Refinements. *Journal of the American Statistical Association,* 83, 687-697 .

Berdai, A. and Garel, B. (1995). Detecting a Univariate Normal Mixture with Two Components. *Statistics and Decisions,* 14, 35-51.

Bernardo, J.M and Giron, F.J. (1988). A Bayesian Analysis of Simple Mixture Problems. In *Bayesian Statistics,* eds. Bernardo J.M., De Groot M.H., Lindley D.V. and Smith A.F.M..

Bertin, E. and Theodoreskou, R. (1995). Preserving Unimodality by Mixing. *Statistics and Probability Letters,* 25, 281-288.

Best, A. and Gipps, B. (1974). An Improved Gamma Approximation to the Negative Binomial. *Technometrics,* 16, 621-624.

Bhattacharya, S.K. (1966). Confluent Hypergeometric Distributions of Discrete and Continuous Type with Application to Accident Proneness. *Bulletin of Calcuta Statistical Association*, 21-31 .

Bhattacharya, S.K. (1967). A Result in Accident Proneness. *Biometrika,* 54, 324-325 .

Bhattacharya, S.K. and Holla, M.S. (1985). On a Discrete Distribution with Special Reference to the Theory of Accident Proneness. *Journal of the American Statistical Association,* 80, 1060-1066.

Binder, D. (1978). Bayesian Cluster Analysis. *Biometrika,* 65, 31-38.

Blischke, W.R (1964). Moment Estimators for the Parameters of a Mixture of Two Binomial Distributions. *Annals of Mathematical Statistics*, 33, 444-454.

Blischke, W.R (1965). Estimating the Parameters Mixtures of the Binomial Distribution. *Journal of the American Statistical Association,* 59, 510-528.

Blum, J.R. and Susarla, V. (1977). Estimation of a Mixing Distribution Function. *Annals of Probability,* 5, 200-209.

Boes, D. (1966). On the Estimation of Mixing Distrbutions. *Annals of Mathematical Statistics,* 37, 177-188.

Boes, D. (1967). Minimax Unbiased Estimator of Mixing Distribution for Finite Mixtures. *Sankhya*, A 29, 417-420.

Bohning, D. (1982). Convergence of Simar's Algorithm for Finding the Maximum Likelihood Estimate of a Compound Poisson Process. *Annals of Statistics,* 10, 1006-1008.

Bohning, D. (1989). Likelihood Inference for Mixtures: Geometrical and Other Constructions of Monotone Step-Length Algorithms. *Biometrika,* 76, 375-383.

Bohning, D. (1995). A Review of Reliable Maximum Likelihood Algorithms for Semiparametric Mixture Models. *Journal of Statistical Planning and Inference,* 47, 5-28.

Bohning, D., Dietz, E., Schaub, R., Schlattman, P. and Lindsay, B. (1994). The Distribution of the Likelihood Ratio for Mixtures of Densities from the One-Parameter Exponential Family. *Annals of the Institute of Statistical Mathematics,* 46, 373-388.

Bohning, D. and Hoffman, K.H. (1982). Numerical Techniques for Estimating Probabilities. *Journal of Statistical Computation and Simulation,* 14, 283-293.

Bohning, D., Schlattman, P. and Lindsay, B. (1992). Computer Assisted Analysis of Mixtures (C.A.M.AN): Statistical Algorithms. *Biometrics,* 48, 283-303.

Bose, S. (1994). Bayesian Robustness with Mixture Classes of Priors. *Annals of Statistics*, 22, 652-667.

Brannas, K. and Rosenqvist, G. (1994). Semiparametric Estimation of Heterogeneous Count Data Models. *European Journal of Operational Research,* 76, 247-258.

Brockett, P. (1977). Approximating Moment Sequences to Obtain Consistent Estimates of Distribution Functions. *Sankhya,* A 39, 32-44.

Brown, S. and Holgate, P. (1971). Tables of the Poisson-Lognormal Distribution. *Sankhya,* B 33, 235-248.

Bryant, J.L. and Paulson, A.S. (1983). Estimation of Mixing Proportions Via Distance Between Characteristic Functions. *Communication in Statistics-Theory and Methods,* 12, 1009-1029.

Bulmer, M.G. (1974). On Fitting the Poisson Lognormal Distribution to Species Abundance Data. *Biometrics,* 30, 101-110.

Burrel, Q. and Cane, V. (1982). The Analysis of Library Data. *Journal of the Royal Statistical Society,* A 145, 439-471.

Cane, V. (1975). The Concept of Accident Proneness. *Bulletin of the Institute of Mathematics,* 15, 183-189.

Cane, V. (1977). A Class of Non-Identifiable Stochastic Models. *Journal of Applied Probability,* 14, 475-482 .

Carlin, B.P. and Lewis, T. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, London.

Carriere, J. (1993). Nonparametric Tests for Mixed Poisson Distributions. *Insurance: Mathematics and Economics,* 12, 3-8.

Cassie, M. (1964). Frequency Distributions Models in the Ecology of Plankton and Other Organisms. *Journal of Animal Ecology,* 31, 65-92.

Celeux, G., Chauveau, D. and Diebolt, J. (1995). On Stochastic Versions of the EM Algorithm. *Technical Report 2514*, *Institute National De Recherche En Informatique Et En Automatique (INRIA).*

Celeux, G. and Diebolt, J. (1985). The SEM Algorithm: a Probabilistic Teacher Algorithm Derived from the EM Algorithm for the Mixture Problem. *Computational Statistics Quarterly,* 2, 73-92.

Celeux, G. and Diebolt, J. (1992). A Stochastic Approximation Type EM Algorithm for the Mixture Problem. *Stochastics and Stochastics Reports,* 41, 119-134.

Chandra, S. (1977). On the Mixtures of Probability Distributions. *Scandinavian Journal of Statistics,* 4, 105-112.

Chatfield, C. and Theobald, M. (1972). Mixtures and Random Sums. *Statistician,* 22, 281-287.

Chauveau, D. (1995). A Stochastic EM Algorithm for Mixtures with Censored Data. *Journal of Statistical Planning and Inference,* 46, 1-25.

Chen, J. (1994). Generalised Likelihood Ratio Test of the Number of Components in Finite Mixture Models. *Canadian Journal of Statistics,* 22, 387-399.

Chen, J. (1995). Optimal Rate of Convergence for Finite Mixture Models. *Annals of Statistics,* 23, 221-233.

Chen, J. and Ahn, H. (1996). Fitting Mixed Poisson Regression Models Using Quasi-Likelihood Methods. *Biometrical Journal,* 38, 81-96.

Chen, J. and Cheng, P. (1994). The Limit Distribution of the Restricted Likelihood Ratio Statistic for Finite Mixture Models. *Technical Report STAT-94-06, University of Waterloo* .

Chen, J. and Cheng, P. (1997). On Testing the Number of Components in Finite Mixture Models with Known Relevant Component Distributions. *Canadian Journal of Statistics,* 25, 389-400.

Chen, J. and Kalbfleisch, J.D. (1996). Penalised Minimum-Distance Estimates in Finite Mixture Models. *Canadian Journal of Statistics,* 24, 167-175.

Chernoff, H. and Lander, E. (1995). Asymptotic Distribution of the Likelihood Ratio Test That a Mixture of Two Binomials is a Single Binomial. *Journal of Statistical Planning and Inference,* 43, 19-40.

Choi, K. (1969). Estimators for the Parameters of Finite Mixtue Distributions. *Annals of the Institute of Statistical Mathematics,* 21, 107-116.

Choi, K. and Bulgren, W. (1968). An Estimation Procedure for Mixtures of Distributions. *Journal of the Royal Statistical Society,* B 30, 444-460.

Clarke, B.R. (1989). An Unbiased Minimum Distance Estimator of the Proportion Parameter in a Mixture of Two Normal Distributions. *Statistics and Probability Letters,* 7, 275-281.

Clarke, B.R. and Heathcote, C.R. (1994). Robust Estimation of k-Components Univariate Mixtures. *Annals of the Institute of Statistical Mathematics,* 46, 83-93.

Clevenson, M.L. and Zidek, J.V. (1975). Simultaneous Estimation of the Means of Independent Poisson Laws. *Journal of the American Statistical Society,* 70, 698-705.

Cohen, A.C. (1967). Estimation in Mixtures of Two Normal Distributions. *Technometrics,* 9, 15-28.

Cox, D. (1983). Some Remarks on Overdispersion. *Biometrika,* 70, 269-274.

Crawford, S. (1994). An Application of Laplace Method to Finite Mixture Distributions. *Journal of the American Statistical Association,* 89, 259-267.

Cressie, N. (1982). A Useful Empirical Bayes Identity. *Annals of Statistics,* 10, 625-629.

Cutler, A. and Cordero-Brana, O. (1996). Minimum Hellinger Distance Estimation for Finite Mixture Models. *Journal of the American Statistical Association,* 91, 1716-1724.

Dalal, S.R. and Hall, W. J. (1983). Approximating Priors by Mixtures of Natural Conjugate Priors. *Journal of the Royal Statistical Society,* B 45, 278-286.

Day, N.E. (1969). Estimating the Components of a Mixture of Normal Distributions. *Biometrika,* 56, 463-474.

De Vaux, R. (1989). Mixtures of Linear Regressions. *Computational Statistics and Data Analysis,* 8, 227-245.

De Vaux, R. and Krieger, A. (1990). Robust Estimation of a Normal Mixture. *Statistics and Probability Letters,* 10, 1-7.

De Vylder, F. (1989). Compound and Mixed Distributions. *Insurance: Mathematics and Economics,* 8, 57-62.

Dean, C.B., Lawless, J. and Willmot, G.E. (1989). A Mixed Poisson-Inverse Gaussian Regression Model. *Canadian Journal of Statistics,* 17, 171-182.

Deely, J. and Kruse, R. (1968). Construction of Sequences Estimating the Mixing Distribution. *Annals of Mathematical Statistics,* 39, 286-288.

Dellaportas, P., Stephens, D.A., Smith, A.F.M. and Cuttman, I. (1995). A comparative study of perinatal mortality using a two-component mixture model. In: *Bayesian Biostatistics*, p. 601-616, Berry D.A. and Stangl D.K. (editors), New York: Marcel Dekker.

Dempster, A.P., Laird, N.M. and Rubin, D. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society,* B 39, 1-38.

Dersimonian, R. (1986). Algorithm AS221 Maximum Likelihood Estimation of a Mixing Distribution. *Applied Statistics,* 35, 302-309.

Dersimonian, R. (1990). Correction to Algorithm AS221. *Applied Statistics,* 39, 176.

Devroye, L. (1992). *Non-Uniform Random Variate Generation*. Springer-Verlag.

Devroye, L. (1993). A Triptych of Discrete Distributions Related to the Stable Law. *Statistics and Probability Letters,* 18, 349-351.

Dick, N.P. and Bowden, D.C. (1973). Maximum Likelihood Estimation for Mixtures of Two Normal Distributions. *Biometrics,* 29, 781-790.

Diebolt, J. and Celeux, G. (1993). Asymptotic Properties of a Stochastic EM Algorithm for Estimating Mixing Proportions. *Communications in Statistics-Stochastic Models,* 9, 599-613.

Diebolt, J. and Ip, E.H.S. (1996). Stochastic EM: Methods and Applications. in *Markov Chain Monte Carlo in Practice*, Eds Gilks W.R., Richardson S., Spiegelhalter D.J., Chapman and Hall.

Diebolt, J. and Robert, C. (1994). Estimation of Finite Mixture Distributions Through Bayesian Sampling. *Journal of the Royal Statistical Society,* B 56, 363-375.

Dietz, E. and Bohning, D. (1994). Analysis of Longitudinal Data Using a Finite Mixture Model. *Statistical Papers,* 35, 203-210.

Douglas, J.B. (1980). *Analysis with Standard Contagious Distributions.* Statistical Distributions in Scientific Work Series 4. International Cooperative Publishing House, Fairland, Maryland USA.

Durairajan, T.M. and Kale, B.K. (1979). Locally Most Powerful Test for the Mixing Proportion. *Sankhya,* 41, 91-100.

Durairajan, T.M. and Kale, B.K. (1982). Locally Most Powerful Similar Test for the Mixing Proportion. *Sankhya,* 44, 153-161.

Eddelman, D. (1988). Estimation of the Mixing Distribution for a Normal Mean with Applications to the Compound Decision Problem. *Annals of Statistics,* 16, 1609-1622 .

Efron, B. and Morris, C. (1975). Data Analysis Using Stein's Estimator and its Generalizations. *Journal of the American Statistical Association,*70, 311-319.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Marcel and Decker.

Engen, S. (1974). On Species Frequency Models. *Biometrika,* 61, 263-270.

Escobar, M.D. and West, M. (1995). Bayesian Density Estimation and Inference Using Mixture. *Journal of American Statistical Association,* 90, 577-588.

Eslinger, P.W.and Woodward, W.A. (1991). Minimum Hellinger Distance Estimation for Normal Models. *Journal of Statistical Computation and Simulation,* 39, 95-113.

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

Everitt, B.S. (1984a). *An Introduction to Latent Variable Models*. Chapman and Hall.

Everitt, B.S. (1984b). Maximum Likelihood Estimation of the Parameters in a Mixture of Two Univariate Normal Distributions: a Comparison of Different Algorithms. *Statistician,* 33, 205-215.

Feller, W. (1943). On a Generalised Class of Contagious Distributions. *Annals of Mathematical Statistics,* 14, 389-400.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*. Vol I, 3rd Edition, Willey New York.

Feng, Z. and McCulloch, C.E. (1992). Statistical Inference Using Maximum Likelihood Estimation and the Generalized Likelihood Ratio when the True Parameter Is on the Boundary of the Parameter Space. *Statistics and Probability Letters,* 13, 325-332.

Feng, Z. and McCulloch, C.E. (1994). On the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture with Unequal Variances. *Biometrics,* 50, 1158-1162.

Feng, Z. and McCulloch, C.E. (1996). Using Bootstrap Likelihood Ratios in Finite Mixture Models. *Journal of the Royal Statistical Society,* B 58, 609-617.

Field, C. and Smith, B. (1994). Robust Estimation- A Weighted Maximum Likelihood Approach. *International Statistical Review,* 62, 405-424.

Finch, S., Mendell, N. and Thode, H. (1989). Probabilistic Measures of Adequacy of a Numerical Search for a Global Maximum. *Journal of the American Statistical Association,* 84, 1020-1023.

Folks, J.L. and Chikara, R.S. (1978). The Inverse Gaussian Distribution and Its Statistical Applications-A Review. *Journal of the Royal Statistical Society,* B 40, 263-289.

Fong, D. and Yip, P. (1993). An EM Algorithm for a Mixture Model of Count Data. *Statistics and Probability Letters,* 17, 53-60.

Fowlkes, E. (1979). Some Methods for Studying the Mixture of Two Normal (Lognormal) Distributions. *Journal of the American Statistical Association,* 74, 561-575.

Fryer, J.G. and Robertson, C.A. (1972). A Comparison of Some Methods for Estimating Mixed Normal Distributions. *Biometrika,* 59, 639-648.

Furman, W.D. and Lindsay, B. (1994a). Testing for the Number of Components in a Mixture of Normal Distributions Using Moment Estimators. *Computational Statistics and Data Analysis,* 17, 473-492.

Furman, W.D. and Lindsay, B. (1994b). Measuring the Relative Effectiveness of Moment Estimators as Starting Values in Maximising Likelihoods. *Computational Statistics and Data Analysis,* 17, 493-508.

Garel, B. (1998). Asymptotic Theory of The Likelihood Ratio Test for the Identification of a Mixture. *Technical Report 05-98, Universite Paul Sabatier, Toulouse*.

Gaver, D. and O'Muircheartaigh, I.G. (1987). Robust Empirical Bayes Analyses of Event Rates. *Technometrics,* 29, 1-15.

Gelfand, A. and Dalal, S. (1990). A Note on Overdispersed Exponential Families. *Biometrika,* 77, 55-64.

Gibbons, R., Clark, D. and Fawcett, J. (1990). A Statistical Method for Evaluating Suicide Clusters and Implementing Cluster Surveilance. *American Journal of Epidemiology, Supplement,* 132, 183-191 .

Goffinet, B. and Loisel, P. (1992). Testing in Normal Mixture Models when the Proportions are Known. *Biometrika,* 79, 842-846.

Good, I.J. (1953). The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrika,* 40, 237-264.

Goutis, C. (1997). Nonparametric Estimation of a Mixing Density via the Kernel Method. *Journal of the American Statistical Association,* 92, 1445-1450.

Grandell, J. (1997). *Mixed Poisson Processes*. Chapman and Hall.

Gray, G. (1994). Bias in Misspecified Mixtures. *Biometrics,* 50, 457-470.

Greenwood, M. and Yule, G. (1920). An Inquiry Into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurence of Multiple Attacks of Disease Or of Repeated Accidents. *Journal of the Royal Statistical Society,* A 83, 255-279.

Guillen, M. and Artis, M. (1992). Count Data Models for a Credit Scoring System. *Presented at the 3rd Meeting of the European Conference Series in Quantitative Economics and Econometrics on Econometrics of Duration, Count and Transition Models, Paris, December, 10-11, 1992.*

Gupta, S. and Huang, W.T. (1981). On Mixtures of Distributions: A Survey and Some New Results On Ranking and Selection. *Sankhya,* B 43, 245-290.

Gurland, J. (1957). Some Interelations Among Compound and Generalised Distributions. *Biometrika,* 44, 263-268.

Gurland, J. (1958). A Generalized Class of Contagious Distributions. *Biometrics,* 14, 229-249.

Gustafson, P. (1996). The Effect of Mixing Distribution Misspesification in Conjugate Mixture Models. *Canadian Journal of Statistics,* 24, 307-318.

Haight, F.A. (1965). On the Effect of Removing Persons with N Or More Accidents from an Accidents Prone Population. *Biometrika,* 52, 298-300.

Hall, P. (1979). On Measures of the Distance of a Mixture from its Parent Distribution. *Stochastic Processes and Applications,* 8, 357-365.

Hall, P. (1981). On the Nonparametric Estimation of Mixture Proportions. *Journal of the Royal Statistical Society,* B 43, 147-156.

Hall, P. and Titterington, D.M. (1984). Efficient Nonparametric Estimation of Mixture Proportions. *Journal of the Royal Statistical Society,* B 46, 465-473 .

Hampel, F., Ronchetti, E., Rousseeuw, P. and Stahel, W. (1986). *Robust Statistics*, Willey and Sons, New York.

Harris, I. (1991). The Estimated Frequency of Zero for a Mixed Poisson Distribution. *Statistics and Probability Letters,* 12, 371-372.

Harris, I. and Basu, A. (1994). Hellinger Distance As a Penalised Loglikelihood. *Communications in Statistics-Simulation and Computation,* 23, 1097-1113.

Hasselblad, V. (1966). Estimation of Parameters for a Mixture of Normal Distributions. *Technometrics,* 3, 431-444.

Hasselblad, V. (1969). Estimation of Finite Mixtures from the Exponential Family. *Journal of the American Statistical Association,* 64, 1459-1471.

Hathaway, R.J. (1985). A Constrained Formulation of Maximum Likelihood Estimation for Normal Mixture Models. *Annals of Statistics,* 13, 795-800.

Hathaway, R.J. (1986a). Another Interpretation of the EM Algorithm for Mixture Distributions. *Statistics and Probability Letters,* 4, 53-56.

Hathaway, R.J. (1986b). A Constrained EM Algorithm for Univariate Normal Mixtures. *Journal of Statistical Computation and Simulation,* 23, 211-230.

Hawkins, R. (1972). A Note On Multiple Solutions to the Mixed Distribution Problem. *Technometrics,* 14, 973-976.

He, X., Simpson, D. and Portnoy, S. (1990). Breakdown Robustness of Tests. *Journal of the American Statistical Association,* 85, 446-452.

Heckman, J. (1990). A Nonparametric Method of Moments Estimator for the Mixture of Exponentials Model and the Mixture of Geometrics Model. in:*Nonparametric and Semiparametric Estimation Models in Econometrics and Statistics,* p. 243-258, Barnettw A, Powell J and Tanchen G(Eds), Cambridge University Press, UK .

Heckman, J., Robb, R. and Walker, J. (1990). Testing the Mixture of Exponentials Hypothesis and Estimating the Mixing Distribution By the Method of Moments . *Journal of the American Statistical Association,* 85, 582-589.

Heckman, J. and Singer, B. (1984). A Method for Minimising the Impact of Distributional Assumptions in Econometric Models for Duration Data. *Econometrica,* 52, 271-320.

Heckman, J. and Walker J. (1990). Estimating Fecundability Data on Waiting Times to First Conseption. *Journal of the American Statistical Association,* 85, 283-294.

Hengartner, N. (1997). Adaptive Demixing in Poisson Mixture Models. *Annals of Statistics*, 25, 917-928.

Henna, J. (1983). A Note on a Consistent Estimator of a Mixing Distribution Function. *Annals of the Institute of Statistical Mathematics,* 35, 229-234.

Henna, J. (1985). On Estimating of the Number of Constituents of a Finite Mixture of Continuous Distributions. *Annals of the Institute of Statistical Mathematics,* 37, 235-240.

Hertier, S. and Ronchetti, E. (1994). Robust Bounded-Influence Tests in General Parametric Models. *Journal of the American Statistical Association,* 89, 897-904.

Hesselager, O. (1994a). A Recursive Procedure for Calculation of Some Compound Distributions. *ASTIN Bulletin,* 24, 19-32.

Hesselager, O. (1994b). Recursions for Certain Bivariate Counting Distributions and Their Compound Distributions. *Working Paper 121, Laboratory of Actuarial Mathematics, University of Copenhagen.*

Hesselager, O. (1996). A Recursive Procedure for Calculation of Some Mixed Compound Poisson Distributions.*Scandinavian Actuarial Journal,* 54-63.

Hinde, J. and Demetrio, C.G.B. (1998). Overdispersion: Models and Estimation. *Computational Statistics and Data Analysis*, 27, 151-170.

Holgate, P. (1970). The Modality of Some Compound Poisson Distributions. *Biometrika,* 57, 666-667.

Holla, M.S. and Bhattacharya, S.K. (1965). On a Discrete Compound Distribution. *Annals of the Institute of Statistical Mathematics,* 15, 377-384.

Hosmer, D. (1973a). A Comparison of Iterative Maximum Likelihood Estimates of the Parameters of a Mixture of two Normal Distributions Under Three Different Types of Sample. *Biometrics,* 29, 761-770.

Hosmer, D. (1973b). On Maximum Likelihood Estimation of the Parameters of a Mixture of Two Normal Distributions When the Sample Size Is Small. *Communications in Statistics,* 1, 217-227.

Irwin, J. (1968). The Generalised Waring Distribution Applied to Accident Theory. *Journal of the Royal Statistical Society,* A 131, 205-225.

Irwin, J. (1975). The Generalised Waring Distribution Parts I,II,III. *Journal of the Royal Statistical Society,* A 138, 18-31,204-227,374-384.

Izenmann, A.J. and Sommer, C. (1988). Philatelic Mixtures and Multimodal Densities. *Journal of the American Statistical Association,* 83, 941-953.

Jamshidian, M. and Jennrich, R. (1997). Acceleration of the EM Algorithm by Using Quasi-Newton Methods. *Journal of the Royal Statistical Society,* B 59, 569-587.

Jansen, M. and Van Duijn, M. (1992). Extensions of Rasch's Multiplicative Poisson Model. *Pshycometrika,* 57, 405-414.

Jewell, N. (1982). Mixtures of Exponential Distributions. *Annals of Statistics,* 10, 479-484.

John, S. (1970). On Analysing Mixed Samples. *Journal of the American Statistical Association,* 65, 755-760.

Johnson, N.L. (1957). Uniqueness of a Result in the Theory of Accident Proneness *Biometrika,* 44, 530-531.

Johnson, N.L. (1967). Note on a Uniqueness of a Result in the Theory of Accident Proneness. *Journal of the American Statistical Association,* 62, 288-289.

Johnson, N.L., Kotz, S.and Kemp, A.W. (1992). *Univariate Discrete Distributions*. 2nd Edition Willey-New York.

Jones, P.N. and McLachlan, G.J. (1992). Improving the Convergence Rate of the EM Algorithm for A Mixture Model Fitted to Grouped Truncated Data. *Journal of Statistical Computation and Simulation,* 43, 31-44.

Jorgensen, M. (1990). Influence-Based Diagnostics for Finite Mixture Models. *Biometrics,* 46, 1047-1058.

Kaas, R. and Hesselager, O. (1995). Ordering Claim Size Distributions and Mixed Poisson Probabilities. *Insurance: Mathematics and Economics,* 17, 193-201.

Kabir, L. (1968). Estimation of Parameters of a Finite Mixture Distributions. *Journal of the Royal Statistical Society,* A 30, 472-482.

Karlis, D. and Xekalaki, E. (1996a). The Power of the Likelihood Ratio Test for the Poisson Distribution. *Technical Report* #23*, Department of Statistics, Athens University of Economics and Business.*

Karlis, D. and Xekalaki, E. (1996b). A Note on the Maximum Likelihood Estimation for Finite Poisson Mixtures. *Technical Report* #24, *Department of Statistics, Athens University of Economics and Busines*s.

Karlis, D. and Xekalaki, E. (1997). Testing for Finite Mixtures Via the Likelihood Ratio Tests. *Technical Report* #27*, Department of Statistics, Athens University of Economics and Business*.

Karlis, D. and Xekalaki, E. (1998a). On Testing for the Number of Components in Finite Poisson Mixture Models. *Annals of the Institute of Statistical Mathematics, (to appear*).

Karlis, D. and Xekalaki, E. (1998b). Minimum Hellinger Distance Estimation for Finite Poisson Mixtures. *Computational Statistics and Data Analysis (to appear).*

Katti, S. (1966). Interelations Among Generalized Distributions and Their Components *Biometrics,* 22, 44-52.

Kazakos, D. (1977). Recursive Estimation of Prior Probabilities Using a Mixture. *IEEE Transactions on Information Theory,* 23, 203-211.

Kemp, A.W. (1986). Weigthed Discrepansies and Maximum Likelihood Estimation for Discrete Distributions. *Communications in Statistics - Theory and Methods,* 15, 783-803.

Kemp, C.D. and Kemp, A.W. (1965). Some Properties of the Hermite Distribution. *Biometrika,* 52, 381-394.

Kemp, C.D. and Kemp, A.W. (1988). Rapid Estimation for Discrete Distributions. *Statistician,* 37, 243-255.

Kemperman, J.H.B. (1991). Mixtures with a Limited Number of Modal Intervals. *Annals of Statistics*, 19, 2120-2144.

Kempton, R.A. (1975). A Generalised Form of Fisher's Logarithmic Series. *Biometrika,* 62, 29-38.

Kempton, R.A. and Taylor, L.R. (1974). Log-Series and Lognormal Parameters as Diversity Discriminants for the Lepidoptera. *Journal of Animal Ecology,* 43, 381-399.

Kling, B. and Goovaerts, M. (1993). A Note on Compound Generalised Distributions. *Scandinavian Actuarial Journal,* 20, 60-72.

Kocherlakota, S. and Kocherlakota, K. (1992). *Bivariate Discrete Distributions*. Marcel and Deccer Inc, New York.

Kumar, K.D., Nicklin, E.H. and Paulson, A.S. (1979). Comment on " Estimating Mixtures of Normal Distributions and Switching Regressions ". *Journal of the American Statistical Association,*74 , 52-56.

Laird, N. (1978). Nonparametric Maximum Likelihood Estimation of a Mixing Distribution. *Journal of the American Statistical Association,* 73, 805-811.

Laird, N. (1982). Empirical Bayes Estimates Using the Nonparametric Maximum Likelihood Estimate for the Prior. *Journal of Statistical Computation and Simulation,* 15, 211-220.

Lambert, D. (1981). Influence Functions for Testing. *Journal of the American Statistical Association,* 76, 649-657.

Lambert, D. (1982). Qualitative Robustness of Tests. *Journal of the American Statistical Association,* 77, 352-357.

Lambert, D. and Tierney, L. (1984). Asymptotic Properties of Maximum Likelihood Estimates in the Mixed Poisson Model. *Annals of Statistics,* 12, 1388-1399.

Lange, K. (1995). A Quasi-Newton Accelaration of the EM algorithm. *Statistica Sinica*, 5, 1-18.

Lavine, M. and West, M. (1992). A Bayesian Method for Classification and Discrimination. *Canadian Journal of Statistics,* 20, 451-461.

Lawless, J. (1987). Negative Binomial and Mixed Poisson Regression. *Canadian Journal of Statistics,* 15, 209-225.

Lemon, G.H. and Krutchkoff, R.G. (1969). An Empirical Bayes Smoothing Technique. *Biometrika,* 56, 361-365.

Leonard, T., Hsu, J., Tsui, K.W. and Murray, J. (1994). Bayesian and Likelihood Inference from Equally Weighted Mixtures. *Annals of the Institute of Statistical Mathematics,* 46, 203-220.

Leroux, B. (1992). Consistent Estimation of a Mixing Distribution. *Annals of Statistics,* 20, 1350-1360.

Leroux, B. and Puterman, M. (1992). Maximum-Penalised-Likelihood for Independent and Markov-Dependent Mixture Models. *Biometrics,* 48, 545-558.

Lesperance, M. and Kalbfleisch, J. (1992). An Algorithm for Computing the Nonparametric MLE of a Mixing Distribution. *Journal of the American Statistical Association,* 87, 120-126.

Lindsay, B. (1981). Properties of the Maximum Likelihood estimator of a mixing distribution. In *Statistical Distributions in Scientific Work* (G.P. Patil, ed), 5, 95-109, Reidel, Boston.

Lindsay, B. (1983a). The Geometry of Mixture Likelihood. A General Theory. *Annals of Statistics,* 11, 86-94.

Lindsay, B. (1983b). The Geometry of Mixture Likelihood. Part II the Exponential Family. *Annals of Statistics,* 11, 783-792.

Lindsay, B. (1986). Exponential Family Mixtures Models with Least Square Estimators. *Annals of Statistics,* 14, 124-137.

Lindsay, B. (1989). Moment Matrices: Application in Mixtures. *Annals of Statistics,* 17, 722-740.

Lindsay, B. (1994). Efficiency Versus Robustness: the Case for Minimum Hellinger Distance and Related Methods. *Annals of Statistics,* 22, 1081-1114.

Lindsay, B. (1995). *Mixture Models: Theory, Geometry and Applications*. Regional Conference Series in Probability and Statistics, Vol 5, Institute of Mathematical Statistics and American Statistical Association.

Lindsay, B. and Basak, P. (1993). Multivariate Normal Mixtures: A Fast Consistent Method of Moments. *Journal of the American Statistical Association,* 88, 468-475 .

Lindsay, B., Clogg, C.C. and Grego, J. (1991). Semiparametric Estimation in the Rasch Model and Related Exponential Response Models, Including a Simple Latent Class for Item Analysis. *Journal of the American Statistical Association,* 86, 96-107.

Lindsay, B. and Roeder, K. (1992). Residuals Diagnostics for Mixture Models. *Journal of the American Statistical Association,* 87, 785-794.

Lindsay, B. and Roeder, K. (1993). Uniqueness of Estimation and Identifiability in Mixture Models. *Canadian Journal of Statistics,* 21, 139-147.

Lindsay, B. and Roeder, K. (1995). A Review of Semiparametric Mixture Models. *Journal of Statistical Planning and Inference,* 47, 29-39.

Lindsay, B. and Roeder, K. (1997). Moment Based Oscillation Properties of Mixture Models. *Annals of Statistics,* 25, 378-386.

Loh, W.L and Zhang, C.H. (1996). Global Properties of Kernel Estimators for Mixing Densities in Exponential Familes Models for Discrete Variables. *Statistica Sinica,* 6, 561-578.

Loh, W.L and Zhang, C.H. (1997). Estimating Mixing Densities in Exponential Family Models for Discrete Variables. *Scandinavian Journal of Statistics,* 24, 15-32.

Louis, T.A. (1982). Finding Observed Information Using the EM Algorithm. *Journal of the Royal Statistical Society,* B 44, 98-130.

Lynch, J. (1988). Mixtures, Generalized Convexity and Balayages. *Scandinavian Journal of Statistics,* 15, 203-210.

Mac Donald, P.D.M. (1971). Comment on "An Estimation Procedure for Mixtures of Distributions" by Choi and Bulgren. *Journal of the Royal Statistical Society,* B 33, 326-329.

Maine, M., Boullion, T. and Rizzuto, G. (1991). Detecting the Number of Components in a Finite Mixture Having Normal Components. *Communications in Statistics- Theory and Methods,* 20, 611-620.

Mallet, A. (1986). A Maximum Likelihood Estimation Method for Random Coefficient Regression Models. *Biometrika,* 73, 645-656.

Maritz, J. (1969). Empirical Bayes Estimation for the Poisson Distribution. *Biometrika,* 56, 349-359.

Maritz, J. and Lwin, T. (1989). *Empirical Bayes* Methods. 2nd Edition.

Markatou, M. (1996). Robust Statistical Inference: Weighted Likelihoods or Usual M-Estimators? *Communications in Statistics-Theory and Methods,* 25, 2597-2613.

Markatou, M. (1998). Mixture Models, Robustness and the Weighted Likelihood Methodology. *Preprint.*

Markatou, M., Basu, A. and Lindsay, B. (1997). Weighted Likelihood Estimating Equations: The Discrete Case with Applications to Logistic Regression. *Journal of Statistical Planning and Inference,* 57, 215-232.

Matusita, K. (1954). On the Estimation by the Minimum Distance Method. *Annals of the Instistute of Statistical Mathematics,* 5, 59-65.

McFadden, J.A. (1965). The Mixed Poisson Proccess. *Sankhya,* A 27, 83-92.

McKay, I. (1996). A Note on Density Estimation for Poisson Mixtures. *Statistics and Probability Letters,* 27, 255-258.

McLachlan, G. (1987). On Bootstraping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture. *Applied Statistics,* 36, 318-324.

McLachlan, G. (1988). On the Choice of Initial Values for the EM Algorithm in Fitting Mixture Models. *Statistician,* 37, 417-425.

McLachlan, G. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Willey Interscience, New York.

McLachlan, G. (1995). On Aitken's Method and Other Approaches for Accelerating Convergence of the EM Algorithm. *Proceedings of the A.C. Aitken Centenary Conference*, Dunedin 1995, 201-209.

McLachlan, G. and Basford, K. (1988). *Mixture Models: Inference and Application to Clustering*. Marcel and Decker Inc.

McLachlan, G. and Jones, P. (1988). Fitting Mixture Models to Grouped and Truncated Data Via the EM Algorithm. *Biometrics,* 44, 571-578.

McLachlan, G. and Krishnan, T. (1997). *The EM Algorithm and Extensions.* Willey Series.

McLachlan, G., Lawoko, C. and Ganesalingam, S. (1982). On the Likelihood Ratio Test for Compound Distributions for Homogeneity of Mixing Proportions. *Technometrics*, 24, 331-335.

McLachlan, G., Mclaren, C.E. and Matthews, D. (1995). An Algorithm for the Likelihood Ratio Test of One Versus Two Components in a Normal Mixture Model Fitted to Grouped and Truncated Data. *Communications in Statistics-Simulation and Computation,* 24, 965-985.

McLaren, C., Brittenham, G. and Hasselblad, V. (1986). Analysis of the Volume of Red Blood Cells: Application of the EM Algorithm to Grouped Data from the Doubly Truncated Lognormal Distribution. *Biometrics,* 42, 143-158.

McNeney, B. and Petkau, J. (1994). Overdispersed Poisson Regression Models for Studies of Air Pollution and Human Health. *Canadian Journal of Statistics,* 22, 421-440.

Meeden, G. (1972). Bayes Estimation of the Mixing Distribution, the Discrete Case. *Annals of Mathematical Statistics,* 43, 1993-1999.

Meilijson, I. (1989). A Fast Improvement of the EM on its Own Terms. *Journal of the Royal Statistical Society,* B 51, 127-138.

Mendell, N., Finch, S.J.and Thode, H.C. (1993). Where is the Likelihood Ratio Test Powerful for Detecting Two Components Normal Mixture? (the Consultant's Forum). *Biometrics,* 49*, 907-915*.

Mendell, N., Thode, H. and Finch, S.J. (1991). The Likelihood Ratio Test for the Two-Component Normal Mixture Problem: Power and Sample Size Analysis. *Biometrics,* 47,1143-1148.

Meng, X.L. and Rubin, D. (1993). Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika,* 80, 267-278.

Meng, X.L. and Van Dyk, D. (1997). The EM Algorithm: An Old Folk Song Sung to a Fast New Tune (with discussion). *Journal of the Royal Statistical Society,* B 59, 511-567.

Mengersen, K. and Robert, C. (1996). Testing for Mixtures: A Bayesian Entropic Approach. *Preprint*.

Milligan, G.W. and Cooper, M.C. (1985). An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Pshychometrika,* 50, 159-179.

Molenaar, W. and Van Zwet, W. (1966). On Mixtures of Distributions. *Annals of Mathematical Statistics,* 37, 201-203.

Morris, C.N. (1982). Natural Exponential Families with Quadratic Variance Functions. *Annals of Statistics,* 10, 65-80.

Murray, G.D. and Titterington, D.M. (1978). Estimation Problems with Data from a Mixture. *Applied Statistics,* 27, 325-334.

Nichols, W.G. and Tsokos, C. (1972). Empirical Bayes Point Estimation in a Family of Probability Distributions. *International Statistical Review,* 40, 147-151.

Nychka, D. (1990). Some Properties of Adding a Smoothing Step to the EM Algorithm. *Statistics and Probability Letters,* 9, 187-193.

Ong, S.H. (1995). Computation of Probabilities of a Generalised Log-Series and Related Distributions. *Communication in Statistics -Theory and Methods,* 24, 253-271.

Ong, S.H. (1996). On a Class of Discrete Distributions Arising from the Birth-Death with Imigration Process. *Metrika,* 43, 221-235.

Ong, S.H. and Muthaloo, S. (1995). A Class of Discrete Distributions Suited to Fitting Very Long Tailed Data. *Communication in Statistics-Simulation and Computation,* 24, 929-945.

Ord, K. (1967). Graphical Methods for a Class of Discrete Distributions. *Journal of the Royal Statistical Society,* A 130 , 232-238.

Ord, K. and Whitmore, G. (1986). The Poisson-Inverse Gaussian Distribution As a Model for Species Abundance. *Communication in Statistics-Theory and Methods,* 15, 853-871.

Oskrochi, G.R. and Davies, R.B (1997). An EM Type Algorithm for Multivariate Mixture Models. *Statistics and Computing,* 7, 145-151.

Ospina, V. and Gerber, H.U. (1987). A Simple Proof of Feller's Characterization of the Compound Poisson Distribution. *Insurance : Mathematics and Economics,* 6, 63-64.

Panaretos, J. (1989). On the Evolution of Surnames. *International Statistical Review,* 57, 161-167.

Panjer, H. (1981). Recursive Evaluation of a Family of Compound Distributions. *ASTIN Bulletin,* 18, 57-68.

Papageorgiou, H. and Wesolowski, J. (1997). Posterior mean identifies the prior distribution in NB and related models. *Statistics and Probability Letters,* 36, 127-134.

Park, C., Basu, A. and Basu, S. (1995). Robust Minimum Distance Inference Based on Combined Distances. *Communication in Statistics-Simulation and Computation,* 24, 653-673.

Parr, W. and Schucany, W. (1982). Minimum Distance Estimation and Components of Goodness-of-Fit Statistics. *Journal of the Royal Statistical Society,* B 44, 178-189.

Patil, G.P. (1964). On Certain Compound Poisson and Compound Binomial Distributions. *Sankhya,* A 27, 293-294.

Patil, G.P. and Rao, C.R. (1978). Weighted Distributions and Size-Biased Sampling with Applications to Widlife Populations and Human Families. *Biometrics,* 34, 179-189.

Patil, G.P., Rao, C.R. and Ratnaparkhi, M.V. (1986). On Discrete Weighted Distributions and Their Use in Model Choise for Observed Data. *Communication in Statistics -Theory and Methods,* 15, 907-918.

Pearson, K. (1894). Contributions to the Mathematical Theory of Evolution. *Philosophical Transanctions of the Royal Society of London,* 185, 71-110.

Perline, R. (1998). Mixed Poisson Distributions Tail Equivalent to Their Mixing Distributions. *Statistics and Probability Letters,* 38, 229-233

Peters, B.C. and Walker, H.F. (1978). An Iterative Procedure for Obtaining Maximum Likelihood Estimates of the Parameters for a Mixture of Normal Distributions. *SIAM Journal of Applied Mathematics,* 35, 362-378.

Pfanzagl, J. (1988). Consistency of Maximum Likelihood Estimators for Certain Nonparametric Families, in Particular: Mixtures. *Journal of Statistical Planning and Inference,* 19, 137-158.

Pfeifer, D. (1987). On the Distance Between Mixed Poisson and Poisson Distributions. *Statistics and Decision,* 5, 367-379.

Phillips, R. (1990). On Constructing Sequences Estimating the Mixing Distribution with Applications. *Communications in Statistics-Simulation and Computation,* 19, 705-720.

Phillipson, C. (1960). The Theory of Confluent Hypergeometric Functions and Its Application to Compound Poisson Proccess. *Scandinavinsk Actuarietidskrift,* 43, 136-162.

Piegorsch, W.W. (1990). Maximum Likelihood Estimation for the Negative Binomial Dispersion Parameter. *Biometrics,* 46, 863-867.

Pielou, E. (1962). Run of One Species with Respect to Another in Transects Through Plant Populations. *Biometrics,* 18, 579-593.

Pilla, R.and Lindsay, B. (1996). Alternative EM Methods in High-Dimensional Finite Mixtures. *Technical Report 96-02, Center for Likelihood Studies, Department of Statistics, Pensylvania State University*.

Polymenis, A. and Titterington, M. (1998). On the Determination of the Number of Components in a Mixture. *Statistics and Probability Letters,* 38, 295-298

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (1992). *Numerical Recipies in FORTRAN: the Art of Scientific Computing*, Cambridge University Press, 2nd Edition.

Preston, P.F. (1971). Estimating the Mixing Distribution by Piece-Wise Polynomial Acrs. *Australian Journal of Statistics,* 13, 64-76.

Quandt, R. and Ramsey, J. (1978). Estimating Mixtures of Normal Distributions and Switching Regressions (with Discussion). *Journal of the American Statistical Association,* 73, 730-752.

Quenouille, M.H. (1949). A Relation Between the Logarithmic,Poisson and Negative Binomial Series. *Biometrics,* 5, 162-164.

Quinn, B., McLachlan, G. and Hjort, N. (1987). A Note On the Aitkin-Rubin Approach to Hypothesis Testing in Mixture Models. *Journal of the Royal Statistical Society,* B 49, 311-314.

Rachev, S.T. and SenGupta, A. (1994). Laplace-Weibull Mixtures for Modeling Price Changes. *Management Science,* 39, 1029-1038.

Rai, G. (1971). A Mathematical Model for Accident Proneness. *Trabajos Estadistica,* 22, 207-212.

Rajagopalan, N. and Loganathan, A. (1991). Bayes Estimates of Mixing Proportions in Finite Mixture Distributions. *Communications in Statistics-Theory and Methods,* 20, 2337-2349.

Razzaghi, M. and Rayens, W. (1987). Modified Maximum Likelihood Estimator for the Mixing Distribution in A Finite Mixture. *Communications in Statistics - Theory and Methods,* 16, 2661-2676.

Read, T. and Cressie, N. (1988). *Goodness of Fit Statistics for Discrete Multivariate Data*. Springer-Verlag.

Redner, R., Hathaway, R. and Bezdek, J. (1987). Estimating the Parameters of Mixture Models with Modal Estimators. *Communications in Statistics- Theory and Methods,* 16, 2639-2660.

Redner, R. and Walker, H. (1984). Mixture Densities, Maximum Likelihood and the EM Algorithm. *SIAM Review,* 26,195-230.

Richardson, S. and Green, P. (1997). On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion). *Journal of the Royal Statistical Society,* B 59, 731-79.

Rider, P. (1961). The Method of Moments Applied to a Mixture of Two Exponential Distributions. *Annals of Mathematical Statistics,* 32, 143-147.

Rider, P. (1962). Estimating the Parameters of Mixed Poisson, Binomial and Weibull Distributions by the Method of Moments. *Bulletin of the International Statistical Institute,* 39, 225-232.

Robbins, H. (1964). The Empirical Bayes Approach to Statistical Desicion Problems. *Annals of Mathematical Statistics,* 35, 1-20.

Robbins, H. (1983). Some Thoughts On Empirical Bayes Estimation. *Annals of Statistics,* 11, 713-723

.

Robert, C.P. (1996). Mixtures of Distributions: Inference and Estimation. in *Markov Chain Monte Carlo in Practice*, Eds Gilks W.R., Richardson S., Spiegelhalter D.J., Chapman and Hall.

Roeder, K. (1994). A Graphical Technique for Determining the Number of Components in a Mixture of Normals. *Journal of the American Statistical Association,* 89, 487-495.

Rolph, J. (1968). Bayesian Estimation of Mixing Distributions. *Annals of Mathematical Statistics,* 39, 1289-1302.

Ross, G.J.S. and Preece, D.A. (1985). The Negative Binomial Distribution. *Statistician,* 34, 323-326.

Rudas, T., Clogg, C.C. and Lindsay, B.G. (1994). A New Index of Fit Based on Mixture Methods for the Analysis Of Contigency Tables. *Journal of the Royal Statistical Society,* B 56, 623-639.

Ruohonen, M. (1988). A Model for the Claim Number Process. *ASTIN Bulletin,* 18, 57-68.

Rutherford, J.R. and Krutchkoff, R.G. (1967). The Empirical Bayes Approach: Estimating the Prior Distribution. *Biometrika,* 54, 326-328.

Sankaran, M. (1969). On Certain Properties of a Class of Compound Poisson Distributions. *Sankhya,* B 32, 353-362.

Sankaran, M. (1970). The Discrete Poisson-Lindley Distribution. *Biometrics,* 26, 145-149.

Sapatinas, T. (1995). Identifiability of Mixtures of Power Series Distributions and Related Characterizations. *Annals of the Institute of Statistical Mathematics,* 47, 447-459.

Scallan, A.J. (1992). Maximum Likelihood Estimation for a Normal - Laplace Mixture Distribution. *Statistician,* 41, 227-231.

Schmidt, P. (1982). An Improved Version of the Quandt-Ramsey MGF Estimator for Mixtures of Normal Distributions and Switching Regressions. *Econometrica,* 50, 501-516.

Schwerder, T. (1982). On the Dispersion of Mixtures. *Scandinavian Journal of Statistics,* 9, 165-169.

Self, S. and Liang, K. (1987). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. *Journal of the American Statistical Association,* 82, 605-610.

Seshadri, V. (1991). Finite Mixtures of Natural Exponential Families. *Canadian Journal of Statistics,* 19, 437-445.

Shaked, M. (1980). On Mixtures from Exponential Families. *Journal of the Royal Statistical Society,* B 42 , 192-198.

Shohat, J. and Tamarkin, J. (1943). *The Problem of Moments*. American Mathematical Society, New York.

Sibuya, M. (1979). Generalised Hypergeometric, Digamma and Trigamma Distributions. *Annals of the Institute of Statistical Mathematics,* 31, 373-390.

Sichel, H.S. (1951). The Estimation of the Parameters of A Negative Binomial Distribution with Special Reference to Psychological Data. *Psychometrica,* 16, 107-127.

Sichel, H.S. (1974). On A Distribution Representing Sentence-Length in Written Prose. *Journal of the Royal Statistical Society,* A 137, 25-34.

Sichel, H.S. (1975). On a Distribution Law for Word Frequencies. *Journal of the American Statistical Association,*70, 542-547.

Sichel, H.S. (1982). Asymptotic Efficiencies of Three Methods of Estimation for the Inverse Gaussian-Poisson Distribution. *Biometrika,* 69, 467-472.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall.

Simar, L. (1976). Maximum Likelihood Estimation of a Compound Poisson Process. *Annals of Statistics,* 4, 1200-1209.

Simon, P. (1955). On a Class of Skew Distributions. *Biometrika,* 42, 425-440.

Simpson, D. (1987). Minimum Hellinger Distance Estimation for the Analysis of Count Data. *Journal of the American Statistical Association,* 82, 802-807.

Simpson, D. (1989). Hellinger Deviance Tests: Efficiency, Breakdown Points and Examples. *Journal of the American Statistical Association,* 84, 107-113.

Smith, A. and Makov, U. (1978). A Quasi-Bayes Procedure for Mixtures. *Journal of the Royal Statistical Society,* B 40, 106-112.

Soromenho, G. (1994). Comparing Approaches for Testing the Number of Components in a Finite Mixture Model. *Computational Statistics,* 9, 65-78.

Sprott, D. (1983). Estimating the Parameters of a Convolution By Maximum Likelihood. *Journal of the American Statistical Association,* 78, 457-460.

Stein, G.Z., Juritz, J. and Zucchini, W. (1987). Parameter Estimation for the Sichel Distribution and Its Multivariate Exetension. *Journal of the American Statistical Association*, 82, 938-944.

Stuart, A. and Ord, K. (1994). *Kendall's Advanced Theory of Statistics, 6th edition, Volume I.* Willey, New York.

Susko, E., Kalbfleisch, J.D. and Chen, J. (1997). Constrained Nonparametric Estimation for Mixture Models. *Technical Report 97-08*, *Department of Statistics and Actuarial Science, Univeristy of Waterloo.*

Symons, M., Grimson, R. and Yuan, Y. (1983). Clustering of Rare Events. *Biometrics,* 39, 193-205.

Tallis, G.M. (1969). The Identifiability of Mixtures of Distributions. *Journal of Applied Probability*, 6 , 389-398.

Tallis, G.M. and Light, R. (1968). the Use of Factorial Moments for Estimating the Parameters of A Mixed Exponential Distribution. *Technometrics,* 10, 161-175.

Tamura, R. and Boos, D. (1986). Minimum Hellinger Distance Estimation for Multivariate Location and Covariance. *Journal of the American Statistical Association,* 81, 223-229.

Tan, W.Y. and Chang, W.C. (1972). Some Comparisons of the Method of Moments and the Method of Maximum Likelihood in Estimating Parameters of A Mixture of Two Normals. *Journal of the American Statistical Association,* 67, 702-708.

Tanner, M. (1991). *Tools for Statistical Inference: Observed Data and Data Augmentation Algorithms*. Lecture Notes in Statistics, Vol 67, Springer- Verlag.

Tanner, M. and Wong, W.H. (1987). the Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association,* 82, 528-551.

Teicher, H. (1961). Identifiability of Mixtures. *Annals of Mathematical Statistics,* 26, 244-248 .

Teicher, H. (1963). Identifiability of Finite Mixtures. *Annals of Mathematical Statistics,* 28, 75-88.

Thode, H., Finch, S. and Mendell, N. (1988). Simulated Percentage Points for the Null Distribution of the Likelihood Ratio Test for a Mixture of Two Normals. *Biometrics,* 44, 1195-1201.

Tierney, L. and Kadane, J. (1986). Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association,* 81, 82-86.

Tierney, L., Kass, R. and Kadane, J. (1989). Fully Exponential Laplace Approximations to Expectations and Variances of Nonpostive Functions. *Journal of the American Statistical Association,* 84, 710-716.

Tierney, L. and Lambert, D. (1984). Asymptotic Efficiency of Estimators of Functionals of Mixed Distributions. *Annals of Statistics,* 12, 1380-1387.

Titterington, D.M, Smith, A.F.M. and Makov, U.E. (1985). *Statistical Analysis of Finite Mixtures Distributions*. Willey and Sons, New York.

Titterington, D.M. (1983). Minimum Distance Non-Parametric Estimation of Mixture Proportions. *Journal of the Royal Statistical Society,* B45, 37-46.

Titterington, D.M. (1990). Some Recent Research in the Analysis of Mixture Distributions. *Statistics,* 21, 619-641.

Tucker, H.G. (1963). An Estimate of the Compounding Distribution of a Compound Poisson Distribution. *Theory of Probability and Applications,* 8, 195-200.

Van De Geer, S. (1995). Asymptotic Normality in Mixture Models. *ESAIM: Probability and Statistics,* 1, 17-33 (Electronic Journal Available in the Address http://www.emath.fr/ps).

Van Der Vaart, A. (1996). Efficient Maximum Likelihood Estimation in Semiparametric Mixture Models. *Annals of Statistics,* 24, 862-878.

Van Houwelingen, J.C. and De Vries, L. (1987). Minimax Estimation of the Mixing Proportion of Two Known Distributions. *Journal of the American Statistical Association,* 82, 300-304.

Vouong, Q.H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica,* 57, 307-333.

Vu, H.T.V. and Maller, R.A. (1996). The Likelihood Ratio Test for the Poisson Versus the Binomial Distribution. *Journal of the American Statistical Association,* 91, 818-825.

Walter, G.G. (1985). Orthogonal Polynomial Estimators of the Prior of A Compound Poisson Distribution. *Sankhya,* A 47, 222-230.

Walter, G.G. and Hamedani, G.G. (1989). Bayes Empirical Bayes Estimation for Discrete Exponential Families. *Annals of the Institute of Statistical Mathematics,* 41, 101-119.

Wang, P., Puterman, M., Cokburn, I. and Le, N. (1996). Mixed Poisson Regression Models with Covariate Dependent Rates. *Biometrics*, 52, 381-400.

Wang, S. and Panjer, H. (1993). Critical Staring Points for Stable Evaluation of Mixed Poisson Probabilities. *Insurance : Mathematics and Economics,* 13, 287-297.

Wang, S. and Sobrero, M. (1994). Further Results On Heselager's Recursive Procedure for Calculation of Some Compound Distributions. *ASTIN Bulletin,* 24, 160-166.

Wang, Y. (1996). Estimation Problems for the Two-parameters Negative Binomial Distribution. *Statistics and Probability Letters,* 26, 113-114.

Wasserman, S. (1983). Distinguishing Between Stochastic Models of Heterogeneity and Contagion. *Journal of Mathematical Phsycology,* 27, 201-215.

Wei, G.C.G. and Tanner, M.A. (1990). A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association,* 85, 699-703.

Weinberg, C. R. and Gladen, B. (1986). The Beta-Geometric Distribution Applied to Comparative Fecundability Studies. *Biometrics,* 42, 547-560.

Whittle, P. (1973). Some General Points in the Theory of Optimal Experimental Designs. *Journal of the Royal Statistical Society,* B 35 , 123-130.

Wilks, S.S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *Annals of Mathematical Statistics,* 9, 60-62.

Willmot, G. (1986). Mixed Compound Poisson Distribution. *ASTIN Bulletin,* 16, Suplement 59-79.

Willmot, G. (1987). The Poisson Inverse Gaussian Distribution as an Alternative to the Negative Binomial. *Scandinavian Actuarial Journal,* 12, 113-127.

Willmot, G. (1989). Asymptotic Tail Behaviour of Poisson Mixtures with Applications. *Advances in Applied Probability,* 22, 147-159.

Willmot, G. (1993). On Recursive Evaluation of Mixed Poisson Probabilities and Related Quantities. *Scandinavian Actuarial Journal,* 18, 114-133.

Willmot, G. and Sundt, B (1989). On Posterior Probabilities and Moments in Mixed Poisson Processes. *Scandinavian Actuarial Journal,* 14, 139-146.

Windham, M. and Cutler, A. (1992). Information Ratios for Validating Mixture Analyses. *Journal of the American Statistical Association,* 87, 1188-1192.

Withers, C.S. (1991). Moment Estimates for Mixtures with Common Scale. *Communications in Statistics- Theory and Methods,* 20, 1445-1461.

Withers, C.S. (1996). Moment Estimates for Mixtures of Several Distributions with Different Means or Scales. *Communications in Statistics- Theory and Methods,* 25, 1799-1824.

Wolfe, J. (1971). Pattern Clustering by Multivariate Mixture Analysis. *Multivariate Behavioral Research,* 5, 329-350.

Woodward, W., Parr, W., Schucany, R. and Lindsey, H. (1984). A Comparison of Minimum Distance and Maximum Likelihood Estimation of a Mixture Proportion. *Journal of the American Statistical Association,* 79, 590-598.

Woodward, W., Whitney, P. and Eslinger, P. (1995). Minimum Hellinger Distance Estimation of Mixture Proportions. *Journal of Statistical Planning and Inference,* 48, 303-319.

Wu, C.F.J. (1983). On the Convergence of the EM Algorithm. *Annals of Statistics,* 11, 95-103.

Xekalaki, E. (1983a). The Univariate Generalised Waring Distribution in Relation to Accident Theory: Proneness, Spells Or Contagion? *Biometrics,* 39, 887-895.

Xekalaki, E. (1983b). Infinite Divisibility, Completeness and Regression Properties of the Univariate Generalised Waring Distribution. *Annals of the Institute of Statistical Mathematics,* 32, 279-289.

Xekalaki, E. (1984a). The Bivariate Generalized Waring Distribution and its Application to Accident Theory. *Journal of the Royal Statistical Society,* A 147, 488-498.

Xekalaki, E. (1984b). Models Leading to the Bivariate Generalized Waring Distribution. *Utilitas Mathimatica*, 35, 263-290.

Xekalaki, E. (1984c). Linear Regression and the Yule Distributions. Journal of *Econometrics,* 24, 397-403.

Xekalaki, E. (1985). Some Identifiability Problems Involving Generalized Waring Distributions. *Publicationes Mathimaticae*, 32, 75-84.

Xekalaki, E. (1987). On an Estimation Procedure for Long Tailed Generalized Waring Distributions. *Proceedings of the First IASC Conference On Computational Statistics and Data Analysis, Tokyo,* 89-96.

Xekalaki, E. and Panaretos, J. (1983). Identifiability of Compound Poisson Distributions. *Scandinavian Actuarial Journal,* 39-45.

Xue, D. and Deddens, J. (1992). Overdispersed Negative Binomial Models. *Communications in Statistics-Theory and Methods,* 21, 2215-2226.

Yiakowitz, S. and Spragins, J. (1969). On the Identifiability of Finite Mixtures. *Annals of Mathematical Statistics,* 39, 209-214.

Ylvisaker, D. (1977). Test Resistance. *Journal of the American Statistical Association,* 72, 551-557.

Zhang, C.H. (1990). Fourier Methods for Estimating Mixing Densities and Distributions. *Annals of Statistics,* 18, 806-831.

Zhang, C.H. (1995). On Estimating Mixing Densities in Discrete Exponential Family Models. *Annals of Statistics,* 23, 929-945.