2023

Finite Mixtures of Mean-Parameterized Conway-Maxwell-Poisson Models

Dongying Zhan
University of Kentucky, dongyingzhan@gmail.com
Digital Object Identifier: https://doi.org/10.13023/etd.2023/469

STUDENT AGREEMENT:

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

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

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

REVIEW, APPROVAL AND ACCEPTANCE

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

Dongying Zhan, Student

Dr. Derek S. Young, Major Professor

Dr. Katherine Thompson, Director of Graduate Studies

Finite Mixtures of Mean-Parameterized Conway-Maxwell-Poisson Models

DISSERTATION

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

By
Dongying Zhan
Lexington, Kentucky

Director: Dr. Derek S. Young, Associate Professor of Statistics
Lexington, Kentucky
2023

ABSTRACT OF DISSERTATION

Finite Mixtures of Mean-Parameterized Conway-Maxwell-Poisson Models

For modeling count data, the Conway-Maxwell-Poisson (CMP) distribution is a popular generalization of the Poisson distribution due to its ability to characterize data over- or under-dispersion. While the classic parameterization of the CMP has been well-studied, its main drawback is that it is does not directly model the mean of the counts. This is mitigated by using a mean-parameterized version of the CMP distribution. In this work, we are concerned with the setting where count data may be comprised of subpopulations, each possibly having varying degrees of data dispersion. Thus, we propose a finite mixture of mean-parameterized CMP distributions. An EM algorithm is constructed to perform maximum likelihood estimation of the model, while bootstrapping is employed to obtain estimated standard errors. A simulation study is used to demonstrate the flexibility of the proposed mixture model relative to mixtures of Poissons and mixtures of negative binomials. An analysis of dog mortality data is presented.

As a generalization of the Poisson distribution and a common alternative to other discrete distributions, the Conway-Maxwell-Poisson (CMP) distribution has the flexibility to explicitly characterize data over- or under-dispersion. The mean-parameterized version of the CMP has received increasing attention in the literature due to its ability to directly model the data mean. When the mean further depends on covariates, then the mean-parameterized CMP regression model can be treated in a generalized linear models framework. In this work, we propose a mixture of mean-parameterized CMP regressions model to apply on data which are potentially comprised of subpopulations with different conditional means and varying degrees of dispersions. An EM algorithm is constructed to find maximum likelihood estimates of the model. A simulation study is performed to test the proposed mixture of mean-parameterized CMP regressions model, and to compare it to model fits using mixtures of Poisson regressions and mixtures of negative binomial regressions. An analysis of the spread of a viral infection in potato plants is performed using these different mixtures of regressions models, where we show the mixture of mean-parameterized CMP regressions to be an effective model.

KEYWORDS: bootstrap, count data, data dispersion, EM algorithm, generalized linear models, negative binomial

Dongying Zhan

December 16, 2023

Finite Mixtures of Mean-Parameterized Conway-Maxwell-Poisson Models

By
Dongying Zhan

Dr. Derek S. Young
Director of Dissertation

Dr. Katherine Thompson
Director of Graduate Studies

December 16, 2023
Date

Dedicated to my courageous, respectful, and beloved sister Scarlett

ACKNOWLEDGMENTS

Firstly, I would like to express my deepest gratitude to my advisor, Dr. Young, for giving me the opportunity to work with him and providing valuable guidance, extreme patience, immense encouragement, and tremendous support to complete my dissertation. I truly appreciated that my advisor diligently worked on my progress and shaped the results into peer-reviewed articles. It was thrilling to debug code together and cooperatively overcome challenges, meanwhile he always shed the light and made a way for me. My advisor serves as a role model, demonstrating not only understanding, prompt responsiveness, and effective communication, but also a tireless dedication to academia, that actually motivated me to keep going throughout my entire study. It's my advisor who made this journey worthwhile, enjoyable, and memorable.

Next, I would also like to express my appreciation to my committee members: Dr. Richard Charnigo, Dr. Anna Smith, Dr. Arnold Stromberg, Dr. Katherine Thompson, Dr. Jianrong Wu, and Dr. Hongbin Zhang, who served on my committee and provided advice on my dissertation. I was grateful to my department for providing a comfortable learning environment and a terrific atmosphere, along with fabulous faculty and staff. My studies in the department not only helped me financially, but also transformed my life by providing ways for me to grow as a professional.

It was very unfortunate that Dr. Stromberg could not serve on my dissertation defense. He was a down-to-earth and humble person, who was always ready to respond and help, especially by walking in another's shoes. Dr. Xiangrong Yin was another unexpected loss to the department during my PhD study. They served as great faculty members with beautiful hearts, intelligent minds, and passionate endeavors

that consistently lighted up the students. How grateful I am that my life ever crossed with them, and the wonderful moments they created will never fade in my memory.

Lastly, I express my gratitude to my family members and friends. My departed mother gave me the greatest love I ever had, although she was not well-educated and even while she suffered from her illness. My father's perseverance, resilience, passion and pursuit in work and life, and strong sense of responsibility for his family have had profound influence on me. I must thank my best buddy, and also my daughter's father, Yu, who greatly helped me in my life with his knowledge, intelligence, generosity, and gentleness in various ways. I feel the essence of life through my daughter, Lucy, for the joy, love, and happiness she has brought me since her birth, as well as her forgiveness of my parenting mistakes. There are numerous friends who are not listed, but who have deeply inspired and enlightened me, with their kindness, sincerity, caring, sharing, and occasionally sophisticated discussions and suggestions. All those experiences continuously strengthen my faith to move forward.

Besides all of the above, I hold a special affection for the city of Lexington, KY. I have lived in this city for more than thirteen years, which is longer than in any other places I have ever lived. It is the place where I have lived a completely new life with a feeling of being accepted and valued. I have many unforgettable memories of knowing many amazing people in this city. The years here make me feel that life is a journey full of adventures, and a better one is always ahead.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

1.1 Finite Mixture Models

1.1.1 Introduction

Finite mixture models constitute a common and broad class of statistical models that employ mathematical approaches to explain various random phenomena. The concept of mixture in a population was noted in the early 19th century when statistics initially emerged as a mathematical tool to study the numbers, primarily from social studies. The earliest idea of mixture model could be traced back to Karl Pearson, one of the pioneers of mathematical statistics, who, in an evolutionary study, considered gender factors within a population, and employed two normal distributions to model fertility (Pearson 1894).

The prominence of finite mixture models increased remarkably when the classic theory of maximum likelihood estimation was applied to parameter estimation. With the rapid advancement of computing technology from the 1980s onward, research on finite mixture models and their applications has been increasingly reported in literature, as evidenced by various books (McLachlan and Peel 2000; Schlattmann 2009; McNicholas 2016) and review articles (Melnykov and Maitra 2010; McLachlan et al. 2019).

The previous and ongoing developments in statistical theory and methodology have provided a robust yet flexible framework, expanding the application of finite mixture models to a wide variety of data structures. Research articles across statistical and scientific fields have demonstrated the effectiveness of finite mixture models to account for population heterogeneity.

The Gaussian mixture model, due to the normal distribution most commonly

assumed in general, was one of the earliest developed (Hasselblad 1966), and remains one of the most widely used mixture models in various fields, including economics (Epps and Epps 1976; Kon 1984), medical studies (Daniel et al. 1991; Bachmann et al. 1999), image analysis (Cox et al. 1996), and many others (Kelly et al. 2009; Brey and Walker 2011). An alternative to Gaussian mixtures is the mixtures of t-distributions, which possess longer tails to accommodate more extreme observations. Multivariate Gaussian mixture model (Biernacki et al. 2003) and multivariate t mixture model (Shoham et al. 2003; Gerogiannis et al. 2009) were also studied for application. The mixture model consisting of exponential distributions was proposed for use in queuing system (Whitt 1984). The mixture of two Weibull distributions has been introduced for modeling the failure data in reliability analysis (Jiang and Murthy 1995).

Finite mixture models in regression settings also play an important role in handling different data structures. The mixture of linear regressions model, was developed early on (Goldfeld and Quandt 1973), and has been extensively studied, particularly for classifying the clustered data (Turner 2000; Viele and Tong 2002). Similar to the generalized linear model extending linear regression, the mixture of generalized linear regressions (Grün and Leisch 2008) allows response variables to be covariates-dependent and assumed to be from non-normal distributions, such as exponential, Poisson, and binomial distributions, among others. R packages, **flexmix** (Leisch 2004) and **mixtools** (Benaglia et al. 2010), are available for mixture modeling on a variety of datasets. The popularity of these packages underscores the appeal of finite mixture models in practice.

1.1.2 Estimating Methods and Related Issues

Estimating the unknown parameters in statistical models is the primary task for statistical analysis, and it can be challenging at times. One of the oldest approaches in mathematical statistics, the method of moments, derives point estimators for statis-

tical models by matching sample moments to their distribution counterparts. Before the era of ubiquitous computing, the method of moments was favored to solve some simple models by hand. For instance, the ordinary least square estimators can be derived for linear regression models using the method of moments. This method was ever used to solve the two components of uni-variate Gaussian mixture model (Pearson 1894). However, the method of moments faces difficulties and accuracy problems when dealing with mixture models which involve large mixtures, as it requires solving a large polynomial system of high-order moment equations. Nevertheless, improved method of moments was reported for solving complex mixture models (Anandkumar et al. 2012; Wu and Yang 2018).

Maximum likelihood estimation is another method for determining the parameter estimates by maximizing the loglikelihood function given the observed data in an assumed model. However, closed-form expressions for maximum likelihood estimators rarely exist in most statistical models. Therefore, numerical optimization techniques are required to find the maximum likelihood estimates. Newton's method provides an iterative algorithm for optimization, but it often encounters issues such as convergence problems, local maxima, or challenges in inverting Hessian matrices, and is seldom used to solve finite mixture models. Regardless, maximum likelihood method has become the dominant method for parameter estimation, since the expectation–maximization (EM) algorithm was introduced (Dempster et al. 1977) and commonly adopted, greatly facilitating the computation for finite mixture models.

While finite mixture models have thrived, several well-known issues in the model setup have been identified as follows:

- **Identifiability** Identifiability is a fundamental property of statistical models. An identifiable model produces distinct parameter estimates from different sets of observations. Poor identifiability can result from model structures and bad data quality, such as insufficient or noisy data, as well as singularity or label switching

issues mentioned below, and substantially may lead to uncertainty or unreliability for statistical modeling. Identifiability for finite mixture models was defined and studied early (Teicher 1963; Yakowitz and Spragins 1968), but it is generally not a concern in the framework of maximum likelihood estimation via the EM algorithm for fitting finite mixture models (McLachlan and Peel 2000).

- **Singularity** Singularity in mathematics may refer to a point that is not differentiable. A matrix is singular if it is not invertible. Singular matrices are common in statistical modeling. For example, multicollinearity can occur in regression modeling when the interdependency among variables makes the correlation and covariance matrices singular. In finite mixture modeling, components may yield singular covariance matrices due to extreme data points or convergence at the boundary of the parameter space, often leading to over-fitting or fitting a meaningless model.

- **Label switching** The label switching problem is caused by the permutation of component labels during the computation in mixture modeling. It is challenging in the context of Bayesian analysis for finite mixture models (Stephens 2000), but is generally not a problem when computing maximum likelihood estimates via the EM algorithm (McLachlan and Peel 2000). A simple strategy to address the label switching problem is to impose identifiable parameter constraints on the estimates so that only one permutation is satisfied in each iteration. Label switching occurs among the bootstrap replicates for calculating the standard errors, but taking the initial values of parameters to start the EM algorithm on each bootstrap sample can completely avoid the problem (McLachlan and Peel 2000).

- **Model Selection** Model selection is crucial for mixture modeling, and primarily involves determining how many components to retain in the mixture. Including unnecessary components in the mixture may result in over-fitting to the data, while including too few components may not reflect the true underlying model. The

loglikelihood-based critera, including but not limited to Akaike's information criterion (AIC; Akaike 1973) and Bayesian information criterion (BIC; Schwarz 1978), along with their modified variants, are prevalent due to their straightforward and easy implementation in the framework of maximum likelihood estimation, and their good performance in model selection. However, the utilization of those criteria for selecting an ideal model is often sensible and depends on the specific situation. The idea of likelihood ratio test (LRT) has been considered to assist in model selection for mixture models (Turner 2000; Chen et al. 2001). However, implementing the LRT for mixture models can be challenging due to the need for effective application of the theoretical development of asymptotic properties to specific mixture models (Dacunha-Castelle and Gassiat 1999).

1.2 Conway-Maxwell-Poisson Distribution

1.2.1 Introduction

The Conway-Maxwell-Poisson distribution (Conway and Maxwell 1961) is named after Richard W. Conway and William L. Maxwell. These two researchers made notable contributions to the field of electrical engineering, and originally developed this distribution to incorporate state-dependent service rates into Poisson queueing model. While the Conway-Maxwell-Poisson distribution may not have the same long-established history as traditional probability distributions, it has gained attention for its ability and flexibility to handle data effectively.

The Conway-Maxwell-Poisson distribution was firstly found useful to fit the tail regions of the purchase data in order to aid retail marketing (Boatwright et al. 2003). The remarkable contribution to this distribution should be attributed to Shmueli et al. (2005), who conducted a systematic investigation into the statistical and probabilistic properties of the Conway–Maxwell–Poisson distribution. This research ignited further studies related to the distribution's applications. The Conway-Maxwell-Poisson

distribution has been adopted in diverse contexts, such as cure rate survival models (Rodrigues et al. 2009) and zero-inflated models (Sellers and Young 2019). Its applications extend to various fields, including transportation data, like motor vehicle crashes (Lord et al. 2008), ecological analysis (Lynch et al. 2014), estimation of COVID-19 mortality (Li and Dey 2022), and more.

As a discrete probability distribution, the Conway–Maxwell–Poisson distribution represents a generalization of the standard Poisson distribution by incorporating an additional parameter. The probability mass function is given by

$$\mathrm{P}(X = x \mid \lambda, \nu) = \frac{\lambda^x}{(x!)^\nu} \frac{1}{\mathcal{Z}(\lambda, \nu)}, \quad x = 0, 1, 2, \ldots, \tag{1.1}$$

where $\mathcal{Z}(\lambda, \nu) = \sum_{x=0}^{\infty} \frac{\lambda^x}{(x!)^\nu}$ is a normalizing constant that guarantees the probabilities of all possible values sum to one. The rate parameter λ, with $\lambda > 0$, is adopted from the Poisson distribution. The dispersion parameter ν, with $\nu \geq 0$, is introduced to adjust the distribution shape relative to the Poisson distribution. It characterizes over-dispersion when $\nu < 1$ (i.e., its variance is greater than its mean), equi-dispersion when $\nu = 1$ (i.e., its variance is equal to its mean), or under-dispersion when $\nu > 1$ (i.e., its variance is less than its mean).

The Conway–Maxwell–Poisson distribution forms a member of the exponential family that share some common and convenient mathematical properties. The sufficient statistics $(\sum_{i=1}^n y_i, \sum_{i=1}^n \log(y_i!))$ for (λ, ν) exist, and a conjugate prior for Bayesian estimation is guaranteed (Shmueli et al. 2005). This distribution encompasses some well-known distributions: Poisson distribution ($\nu = 1$), geometric distribution ($\nu = 0, 0 < \lambda < 1$), and Bernoulli distribution ($\nu \to \infty$). For a comprehensive overview of the Conway-Maxwell-Poisson distribution, including its statistical theory, methodology, and practical models, the most recent monograph by Dr. Kimberly Sellers (2023) serves as a valuable resource.

1.2.2 Parameterization

Parametric distributions can be characterized by various parameter sets. For example, the gamma distribution features shape/rate parameters or alternative shape/scale parameters, both of which are commonly used. Parameterizing a distribution through mathematical manipulation doesn't change the probabilities or likelihood in the context of maximum likelihood estimation, but it usually offers the advantages of convenience and usefulness.

The negative binomial distribution has been presented with various parameterizations. Originally, the negative binomial random variable was defined based on independent Bernoulli trials to count the number of failures until a fixed number of successes being achieved. This led to parameterization based on the number of successes and the probability of success in each trial. Another parameterization emerged by deriving negative binomial as a Poisson-gamma mixture, where the Poisson mean parameter, instead of a constant assumed, but follows a gamma distribution with its own mean as 1. This parameterization of negative binomial is characterized by a location parameter or distribution mean from the Poisson distribution and a shape parameter from the gamma distribution (Hilbe 2011). This new configuration makes negative binomial distribution more convenient for practical applications.

Similarly, the parameterization approach has been successfully applied to the Conway–Maxwell–Poisson distribution. The rate parameter λ initially adopted from Poisson distribution loses its ability to indicate the distribution's location as the distribution mean. This limitation hindered the distribution's application due to its frustrating aspect. Given that the Conway–Maxwell–Poisson distribution doesn't have closed-form expressions for its mean and variance in terms of its parameters, Shmueli et al. (2005) provided the approximations of the mean $\mathrm{E}[Y] \approx \lambda^{1/\nu} + \frac{1}{2\nu} - \frac{1}{2}$ and the variance $\mathrm{Var}[Y] \approx \frac{1}{\nu}\lambda^{1/\nu}$ by using the asymptotic expression for the normalizing constant $\mathcal{Z}(\lambda, \nu)$. These approximations have been shown to be accurate while $\nu \leq 1$

and $\lambda > 10^{\nu}$ (Shmueli et al. 2005), making these approximations particularly useful for over-dispersion and equi-dispersion. Guikema and Coffelt (2008) took $\mu_{\star} = \lambda^{1/\nu}$ to reparameterize the center of the Conway-Maxwell-Poisson distribution, a method adopted by SAS/ETS COUNTREG procedures (SAS Institute Inc. 2013). Ribeiro et al. (2020) used the approximation of the distribution mean, $\mu \approx \lambda^{1/\nu} + \frac{1}{2\nu} - \frac{1}{2}$, to locate the center of the Conway–Maxwell–Poisson distribution for a regression model.

The previously mentioned parameterizations rely on the approximation of the distribution mean, rendering their results accurate only for part of the parameter space. By contrast, Huang (2017) reparameterized the Conway–Maxwell–Poisson distribution via the distribution's true mean. This approach results in a mean-parameterized Conway–Maxwell–Poisson distribution, parameterized by the mean parameter and dispersion parameter. The true-mean parameterization is considered valid over the entire parameter space, outperforming previous approximations (Guikema and Coffelt 2008; Ribeiro et al. 2020). Moreover, the mean-parameterized Conway–Maxwell–Poisson distribution exhibits a satisfying property that different values of dispersion parameter appear comparable for the same mean value (Huang 2017). This parameterization proves to be an useful approach for meaningful and practical applications of the Conway–Maxwell–Poisson distribution.

1.2.3 Related Distributions

The Conway–Maxwell–Poisson distribution is a valuable addition to the family of discrete distributions. While no single distribution can guarantee its performance outperforming the others in statistical modeling, it is necessary to consider some related discrete distributions.

The Poisson distribution, a classical statistical model, is the most extensively studied and applied model for count data. It is commonly used to model the probability of a specific number of occurrences. The Poisson random variable typically

describes the rare events, assuming that the occurrences of the event are independent, non-simultaneous, and occur at a constant rate over time (represented by the mean rate). Notably, the Poisson distribution assumes equi-dispersion (i.e., the mean and variance are equal), that often doesn't hold in real-world data.

Over-dispersion, characterized by greater variability in data than what is allowed based on the model assumed, has been a long-standing issue in discrete modeling (Cox 1983; Hinde and Demétrio 1998; Dean and Luncy 2016). General modeling sometimes leads to a lack of fit due to over-dispersion. Over-dispersion can result from various sources, such as assumption failures involving dependency between responses, missing covariates, extreme outliers, and zero-inflation, as well as other possible causes. Consequently, more accurate models are needed to address this issue. The negative binomial distribution can be viewed as an extension of Poisson distribution, since it is alternatively derived as a Poisson–gamma mixture aforementioned in Sect. 1.2.2, and particularly it exhibits a variance greater than its mean. It is not surprising that the negative binomial distribution is better qualified to handle over-dispersion than the Poisson distribution, and it has become a convention for modeling over-dispersed data.

There are numerous generalizations of Poisson distribution available. These generalizations usually contain the Poisson distribution as a special case, and introduce new parameters to cope with both over-dispersion and under-dispersion. The generalized Poisson distribution gained some popularity (Consul and Jain 1973; Consul 1989). The weighted Poisson distribution was proposed by Castillo and Pérez-Casany (1998) and developed as a weighted Poisson mixture. The hyper-Poisson distribution was constructed as a power series distribution and is a subclass of the three-parameter hypergeometric series distribution (Bardwell and Crow 1964). Unlike the original Poisson distribution, which has a solid probabilistic foundation to characterize Poisson process, these generalizations often seek a desired fit empirically without

phenomenon-based understanding, resulting in a lack of intuitive interpretation of their parameters.

However, the Conway-Maxwell-Poisson distribution, particularly the mean-parameterized Conway-Maxwell-Poisson distribution, stands out among these generalizations as a proximal alternative to the Poisson distribution. Its simple yet effective modification, based on Poisson distribution, enables it to model both over-dispersion and under-dispersion.

1.3 Overview of the Dissertation

The rest of the dissertation is organized as follows:

- Chapter 2 is the paper titled "Finite Mixtures of Mean-Parameterized Conway–Maxwell–Poisson Models," which is currently in press with *Statistical Papers*.

- Chapter 3 is the paper titled "Finite Mixtures of Mean-Parameterized Conway-Maxwell-Poisson Regressions," initially submitted to the *Journal of Statistical Theory and Practice*. This paper is currently under review after a major revision, with improvements primarily attributed to the essential contributions of my advisor. These improvements are discussed in Chapter 4.

- Chapters 2 and 3 propose finite mixtures of mean-parameterized Conway-Maxwell-Poisson (regressions) models. Chapter 2 focuses on the univariate setting, while Chapter 3 explores the regression setting on the means. In both chapters, the EM algorithm was utilized for maximum likelihood estimation, and the parameters were estimated for the mixtures of two or three mean-parameterized Conway-Maxwell-Poisson components in the simulation study.

- Moreover, Poisson (regression) mixtures and negative binomial (regression) mixtures models were used for model selection in Chapter 2, and model comparison in

Chapter 3. The dog mortality data in Chapter 2, and the aphids data in Chapter 3 were applied, respectively, to the corresponding models.

- Chapter 4 provides conclusions and some discussions for the study. Improvements for Chapter 3 are available and applied to SIDS data in Chapter 4. Additionally, future research to regress on mixing proportions in the mixture of mean-parameterized Conway-Maxwell-Poissson regressions model and some considerations regarding singularity are included in Chapter 4.

Chapter 2 Finite Mixtures of Mean-Parameterized Conway-Maxwell-Poisson Models

2.1 Introduction

Count data are ubiquitous in many important applications, such as when recording the number of insurance claims filed by individual policyholders (Smyth and Jørgensen 2002; Yip and Yau 2005), the number of car crashes on a particular road segment (Abdel-Aty and Essam Radwan 2000; Konşuk Ünlü et al. 2022), the number of adverse events or deaths in biomedical research (Zhang et al. 2018; Muenz et al. 2018), or the number of a particular species of wildlife in nature (Cunningham and Lindenmayer 2005; Dénes et al. 2015). Observed counts from such settings often come from a relatively small subset of the natural numbers $\mathbb{N}$, for which the Poisson or negative binomial distributions tend to be good-fitting models.

The Poisson distribution is parameterized by the rate parameter $\lambda > 0$, which is the average number of events that occur. The Poisson distribution characterizes data that are equi-dispersed; i.e., the variance is equal to the mean. This is often unrealistic in most observed count datasets. Departures from the equi-dispersion assumption of Poisson models include where the data are over-dispersed (i.e., the variance of the data is *greater* than what is expected under the Poisson model, which is the mean) or under-dispersed (i.e., the variance of the data is *less* than what is expected under the Poisson model, which is, again, the mean). The degree of dispersion in a dataset is often measured by the dispersion index, which is the ratio of the data's variance to its mean.

The negative binomial distribution is, perhaps, the most commonly-employed discrete distribution used to model over-dispersed count data. In classic probability theory, a negative binomial variable is built from independent Bernoulli trials to count

the number of failures until a fixed number of successes is achieved. This is what Hilbe (2011) calls the *type I negative binomial (NB1) distribution*. Alternatively, the negative binomial distribution can be derived using a Poisson-gamma mixture distribution. In this latter setting, the negative binomial distribution is parameterized by the mean parameter $\mu > 0$ and the gamma shape parameter $\alpha > 0$, thus yielding the variance of the negative binomial distribution as $\mu + \mu^2/\alpha$. Hilbe (2011) calls this parameterization the *type II negative binomial (NB2) distribution*. It is clear that this form of the negative binomial distribution can characterize over-dispersion relative to the Poisson assumption since the variance is greater than the mean.

The Conway-Maxwell-Poisson (CMP) distribution is another popular discrete distribution that generalizes the Poisson distribution by incorporating a dispersion parameter to handle both over-dispersion and under-dispersion. The distribution was originally proposed in Conway and Maxwell (1961) to model queueing systems with state-dependent service rates. While used occasionally in the literature thereafter, the current popularity of the distribution can be attributed to Shmueli et al. (2005), who studied the statistical and probabilistic properties of the CMP distribution as well as presented various estimation methods. They further demonstrated the CMP's utility by modeling two datasets: one on the sales of a particular item of clothing from a large national retailer and one about the lengths of words in a Hungarian dictionary. Since then, various novel model extensions using the CMP distribution have been used for many interesting applications, including the development of CMP regression in a generalized linear models (GLM) framework for a refined analysis of motor vehicle crashes (Lord et al. 2008) and the modelling of airfreight breakages (Sellers and Shmueli 2010), a spatio-temporal CMP model for estimating COVID-19 mortality data in the United States (Li and Dey 2022), and a multivariate CMP model for analyzing the numbers of goals scored by the home and away teams in the English Premier League from 2018 to 2021 (Piancastelli et al. 2022). We refer

the reader to the recent monograph of Sellers (2023) for a modern and expansive treatment of CMP models. Note that the acronyms used for various CMP models in that monograph are employed in the present work.

Other extensions to CMP models have been developed to better characterize the perceived amount of dispersion in the data. For example, Sellers and Shmueli (2013) use a log link function applied to the CMP dispersion parameter such that the linear predictor in this relationship includes dummy variables to model varying group-level dispersions. Sellers and Raim (2016) address the increased chance of data over-dispersion due to excess zeroes through a zero-inflated CMP (ZICMP) regression model. Arora et al. (2021) develop a zero- and k-inflated CMP (ZkICMP) model for further flexibility in characterizing data dispersion when inflation occurs at both 0 and another positive count $k \in \mathbb{N}_+$. One setting where there is scant treatment of CMP distributions is when count data for a particular application may be comprised of subpopulations, each possibly having varying degrees of dispersion, but the subpopulation to which an individual observation belongs is unobserved. Specifically, this would be a finite mixture model where the components are CMP distributions. Sur et al. (2015) is the only work to our knowledge that addresses such a mixture model. However, that paper is focused on a very specific setting where the data is observed in the range of $t, t+1, \ldots, T$, so that a truncated CMP distribution (truncated at values below t and above T) is considered. Moreover, the authors are attempting to characterize bimodality that results from peaks at the values of t and T, so the resulting model is a two-component mixture of truncated CMPs. This model is demonstrated to be effective in characterizing the number of days spent in a hospital (where $t = 1$ and the counts are censored at $T = 15$) and to characterize data on a Likert scale (one example involves a marketing survey research question and another involves online ratings of a particular hotel).

Our work develops a general m-component mixture of CMPs where m is allowed

to be greater than 2. However, we do not consider a truncated version of the CMP, but rather we use a mean-parameterized version of the CMP for our component distributions. This has the benefit of yielding component means that not only have a more intuitive interpretation, but also can be easily compared with component means from, say, mixtures of Poissons or mixtures of negative binomials. We further address other considerations, such as the estimation of standard errors for the parameters in our model and determining the number of components through model selection criteria, neither of which were addressed in the model presented in Sur et al. (2015). We perform maximum likelihood estimation using an expectation-maximization (EM) algorithm (Dempster et al. 1977), and show that the model is informative for a dog mortality dataset.

The rest of this paper is organized as follows. In Sect. 2.2, we provide a brief review of both the classic CMP and mean-parameterized CMP distributions. In Sect. 2.3, we have a general discussion about finite mixture models where the components are discrete distributions. Emphasis will be given to the Poisson, negative binomial, and mean-parameterized CMP settings. In Sect. 2.4, we present details for maximum likelihood estimation of mixture of mean-parameterized CMP models via an EM algorithm. In Sect. 2.5, we conduct a simulation study on the performance of our model, including an assessment of model selection criteria for determining the number of components and comparing to other mixtures of discrete distributions. In Sect. 2.6 we analyze a dog mortality dataset. We conclude with some final comments in Sect. 2.7.

2.2 Mean-Parameterized CMP Distribution

The flexible queueing system considered in Conway and Maxwell (1961) is a single-queue, single-server system with random arrival times and a first-come-first-serve policy for arriving units, where the interarrival times and service times are each

exponentially distributed. Under these assumptions, the authors define a system of differential equations with an assumed steady state and further define the quantity λ to be the product of the mean of the exponential distribution for the interrarival times and the mean service for a unit when it is the only unit in the system. This results in a set of recursion equations (that depend on λ), which upon solving yield the CMP distribution. The classic CMP distribution is a member of the exponential family, and is characterized by the rate parameter $\lambda > 0$ and dispersion parameter $\nu \geq 0$. The probability mass function (pmf) for a CMP random variable Y is given by

$$\mathrm{P}(Y = y \mid \lambda, \nu) = \frac{\lambda^y}{(y!)^\nu} \frac{1}{\mathcal{Z}(\lambda, \nu)}, \quad y = 0, 1, 2, \ldots, \tag{2.1}$$

where $\mathcal{Z}(\lambda, \nu) = \sum_{y=0}^{\infty} \frac{\lambda^y}{(y!)^\nu}$ is a normalizing constant that guarantees the pmf sums to unity. For such a random variable Y, $\lambda = \mathrm{E}[Y^\nu]$ is a generalized form of the Poisson rate parameter, and ν, which does not appear in the Poisson distribution, allows for adjustment of the rate of decay. Note that if you have a sample of n *iid* values, say $y_1, \ldots, y_n$, from the above CMP distribution, the factorization theorem can be applied to the corresponding loglikelihood to show that $(\sum_{i=1}^n y_i, \sum_{i=1}^n \log(y_i!))$ is sufficient for (λ, ν) (Shmueli et al. 2005).

The flexibility of the CMP pmf in (2.1) is attributed to how it generalizes the pmfs for a number of other discrete distributions. Specifically, we have a Poisson distribution with rate λ when $\nu = 1$, a geometric distribution with success probability $(1 - \lambda)$ when $\nu = 0$ and $\lambda < 1$, and a Bernoulli distribution with success probability $\lambda/(1+\lambda)$ as $\nu \to \infty$. Moreover, the dispersion parameter in the classic CMP provides the ability to indicate over-dispersion when $\nu < 1$ and under-dispersion when $\nu > 1$. However, unlike the rate parameter λ in a Poisson distribution and the mean parameter μ in a negative binomial distribution, the parameter λ in the CMP distribution does not directly indicate the center of the distribution nor explain the occurrence

rate in the data. Moreover, the CMP distribution does not have closed-forms for its mean and variance in terms of λ and ν, which again obfuscates the interpretations of estimates for those parameters. Regardless, Shmueli et al. (2005) utilize the asymptotic expression of the normalizing constant $\mathcal{Z}$ to provide approximations for the mean and variance. These are given by, respectively,

$$\mathrm{E}[Y] \approx \lambda^{1/\nu} + \frac{1}{2\nu} - \frac{1}{2} \quad \text{and} \tag{2.2}$$

$$\mathrm{Var}[Y] \approx \frac{1}{\nu}\lambda^{1/\nu}, \tag{2.3}$$

which are considered accurate while $\nu \leq 1$ and $\lambda^{1/\nu} > 10$.

Some authors worked on reparameterizing the CMP distribution in (2.1) to fit within a GLM-type framework. The approximated CMP (ACMP) of Guikema and Coffelt (2008) took $\mu_\star = \lambda^{1/\nu}$ to reparameterize the center of the CMP distribution. This method was adopted to use within SAS/ETS COUNTREG procedures. Ribiero et al. (2020) presented a mean-parameterized CMP (MCMP2) model that used the approximation of the distribution mean $\mu \approx \lambda^{1/\nu} - \frac{\nu-1}{2\nu}$ to locate the center of the CMP distribution. Since these reparameterizations rely on the approximation of the normalizing constant $\mathcal{Z}$, the results are only accurate for part of the parameter space.

Huang (2017) also reparameterized the CMP distribution via the distribution mean μ. This mean-parameterized CMP (MCMP1) distribution can be characterized by the mean parameter $\mu \geq 0$ and dispersion parameter $\nu \geq 0$ such that its pmf is

$$\mathrm{P}(Y = y \mid \mu, \nu) = \frac{\lambda(\mu,\nu)^y}{(y!)^\nu} \frac{1}{\mathcal{Z}(\lambda(\mu,\nu),\nu)}, \quad y = 0, 1, 2, \ldots, \tag{2.4}$$

where the rate parameter $\lambda(\mu, \nu)$ can be solved as a function of μ and ν by using

$$\sum_{y=0}^{\infty} (y - \mu) \frac{\lambda(\mu,\nu)^y}{(y!)^\nu} = 0. \tag{2.5}$$

This reparameterization via the mean is found by solving the nonlinear equation (2.5) instead of adopting an approximation as done in the MCMP2 model. The MCMP1 distribution has other appealing properties. For example, μ and ν are orthogonal (Huang and Rathouz 2017), while the values of ν appear comparable for the same value of μ (Huang 2017). Therefore, the MCMP1 provides, perhaps, a more ideal approach when seeking meaningful and practical interpretations from an estimated CMP model.

2.3 Finite Mixtures of Discrete Distributions

The random variable Y follows an m-component (parametric) mixture distribution if it has the mixture density

$$g(y; \boldsymbol{\Psi}) = \sum_{j=1}^{m} \pi_j f_j(y; \boldsymbol{\theta}_j), \tag{2.6}$$

where the π_js are mixing proportions that satisfy $0 \leq \pi_j \leq 1$, $j = 1, \ldots, m$ and $\sum_{j=1}^{m} \pi_j = 1$. Depending on whether Y is a continuous or discrete random variable, the f_j are, respectively, component-specific density or mass functions from a parametric family with $\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_j \subseteq \mathbb{R}^q$, where $\boldsymbol{\Theta}_j$ is open in $\mathbb{R}^q$. The mixture density in (2.6) is then parameterized by $\boldsymbol{\Psi} = \left(\pi_1, \ldots, \pi_{m-1}, \boldsymbol{\theta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_m^{\mathrm{T}}\right)^{\mathrm{T}}$. In the present work, we focus on finite mixtures of discrete distributions, so $Y \in \mathbb{N}$. For the majority of applications, including the dog mortality data analyzed later, the observed (count) values are typically not large. Moreover, we focus on mixtures of Poissons, mixtures of negative binomials, and mixtures of MCMP1, which are now defined in turn.

The m-component mixture of Poissons model has the mixture density

$$g(y; \boldsymbol{\Psi}) = \sum_{j=1}^{m} \pi_j \frac{e^{-\lambda_j} \lambda_j^y}{y!}, \tag{2.7}$$

where the parameter vector is $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{m-1}, \lambda_1, \ldots, \lambda_m)^{\mathrm{T}}$. Compared to a single Poisson distribution, mixtures of Poissons can characterize over-dispersion (relative to a single Poisson distribution) by including more equi-dispersed components. Two- and three-component mixtures of Poissons were found to provide good fits for insurance claims data (Ismail et al. 2004). Mixtures of Poissons were also applied to identify different RNA polymerase II distributions in the transcribed regions (Feng et al. 2008) and to inform document classification (Li and Zha 2006).

The m-component mixture of negative binomials model has the mixture density

$$g(y; \boldsymbol{\Psi}) = \sum_{j=1}^{m} \pi_j \frac{\Gamma(y + \alpha_j)}{y!\Gamma(\alpha_j)} \left(\frac{\alpha_j}{\mu_j + \alpha_j}\right)^{\alpha_j} \left(\frac{\mu_j}{\mu_j + \alpha_j}\right)^{y}, \tag{2.8}$$

where the parameter vector is $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{m-1}, \mu_1, \ldots, \mu_m, \alpha_1, \ldots, \alpha_m)^{\mathrm{T}}$. The μ_js and the α_js are, respectively, the mean parameters and dispersion parameters for the components. Mixtures of negative binomials have been used to model vehicle crash data (Park and Lord 2009; Zou et al. 2013) and the frequency of earthquakes (Huang et al. 2019). Likewise, mixtures of negative binomials have also been found as an effective model-based clustering tool in RNA-seq count studies (Li et al. 2018).

Finally, we define our proposed m-component mixture of MCMP1 model. Using the definition we supplied in Sect. 2.2, this model has the mixture density

$$g(y; \boldsymbol{\Psi}) = \sum_{j=1}^{m} \pi_j \frac{\lambda(\mu_j, \nu_j)^y}{(y!)^{\nu_j}} \frac{1}{\mathcal{Z}(\lambda(\mu_j, \nu_j), \nu_j)}, \tag{2.9}$$

where the parameter vector is $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{m-1}, \mu_1, \ldots, \mu_m, \nu_1, \ldots, \nu_m)^{\mathrm{T}}$. Similar to the mixture of negative binomials, the μ_js are mean parameters and the ν_js are dispersion parameters for the components. Beyond the advantages of interpretability for each component's MCMP1 distribution, we are also able to leverage some of the estimation tools that are presented in Huang (2017). These are incorporated in the

next section and Appendix A.

2.4 EM Algorithm for Maximum Likelihood Estimation

In this section, we provide the details to perform maximum likelihood estimation of the mixture of MCMP1 model using an EM algorithm. Let $Y_i \in \mathbb{N}$, $i = 1, \ldots, n$ be iid random variables following the m-component mixture model in (2.9) parameterized by $\boldsymbol{\Psi}$ which was defined immediately following the density. Let $\boldsymbol{y} = (y_1, \ldots, y_n)^\intercal$ be a vector of the corresponding realizations of the Y_is. The likelihood function for this model is then given by

$$\mathcal{L}_o(\boldsymbol{\Psi}; \boldsymbol{y}) = \prod_{i=1}^{n} \prod_{j=1}^{m} \pi_j \frac{\lambda(\mu_j, \nu_j)^{y_i}}{(y_i!)^{\nu_j}} \frac{1}{\mathcal{Z}(\lambda(\mu_j, \nu_j), \nu_j)}. \tag{2.10}$$

The corresponding loglikelihood function is then given by

$$\ell_o(\boldsymbol{\Psi}; \boldsymbol{y}) = \sum_{i=1}^{n} \sum_{j=1}^{m} \left\{ \log \pi_j + \log \left(\frac{\lambda(\mu_j, \nu_j)^{y_i}}{(y_i!)^{\nu_j}} \frac{1}{\mathcal{Z}(\lambda(\mu_j, \nu_j), \nu_j)} \right) \right\}. \tag{2.11}$$

Note that we use the "o" subscript on both of the above functions to indicate that these are the likelihood and loglikelihood functions based on the observed data.

As is typical in maximum likelihood estimation of mixture models, the likelihood in (2.10) is difficult to directly optimize. To make optimization tractable, we begin by noting that the observations $\boldsymbol{y}$ are considered incomplete, because their corresponding component labels are not observed. To make the data complete, we define the indicator variables $Z_{ij} \sim \text{Bern}(\pi_j)$ to be the (unobserved) component label for observation i; specifically, $Z_{ij} = \mathtt{I}\{\text{if } Y_i \text{ is from component } j\}$. Letting $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{im})^\intercal$, it follows that

$$\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n \overset{iid}{\sim} \text{Mult}_m(1, \{\pi_1, ..., \pi_m\}), \tag{2.12}$$

where $\text{Mult}_m(\cdot, \cdot)$ denotes the multinomial distribution with m categories.

The complete data are then given by $(y_1, \boldsymbol{Z}_1^\intercal), \ldots, (y_n, \boldsymbol{Z}_n^\intercal)$, which are then used to form the complete-data likelihood

$$\mathcal{L}_c(\boldsymbol{\Psi}; \boldsymbol{y}, \boldsymbol{Z}) = \prod_{i=1}^{n} \prod_{j=1}^{m} \left\{ \pi_j \frac{\lambda(\mu_j, \nu_j)^{y_i}}{(y_i!)^{\nu_j}} \frac{1}{\mathcal{Z}(\lambda(\mu_j, \nu_j), \nu_j)} \right\}^{Z_{ij}}. \tag{2.13}$$

Moreover, the complete-data loglikelihood is given by

$$\ell_c(\boldsymbol{\Psi}; \boldsymbol{y}, \boldsymbol{Z}) = \sum_{i=1}^{n} \sum_{j=1}^{m} Z_{ij} \left\{ \log \pi_j + \log \left(\frac{\lambda(\mu_j, \nu_j)^{y_i}}{(y_i!)^{\nu_j}} \frac{1}{\mathcal{Z}(\lambda(\mu_j, \nu_j), \nu_j)} \right) \right\}. \tag{2.14}$$

The "c" subscript used on the above functions is to indicate that these are the likelihood and loglikelihood functions based on the complete data. As noted earlier, the $\boldsymbol{Z}_i$s are unobserved, and thus treated as missing in the loglikelihood function (2.14). Using the above complete-data setup, we can turn to perform maximum likelihood estimation via an EM algorithm.

E-Step Given a fixed $\boldsymbol{\Psi}^{(t)}$ at the t^{th} iteration, $t = 0, 1, \ldots$, the conditional expectation of $\ell_c(\boldsymbol{\Psi}; \boldsymbol{y}, \boldsymbol{Z})$ given the observed data $\boldsymbol{Y} = \boldsymbol{y}$ is computed as

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(t)}) = \mathrm{E}_{\boldsymbol{\Psi}^{(t)}}[\ell_c(\boldsymbol{\Psi}; \boldsymbol{y}, \boldsymbol{Z})] \tag{2.15}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} z_{ij}^{(t)} \Big\{ \log \pi_j + \log \Big(\frac{\lambda(\mu_j, \nu_j)^{y_i}}{(y_i!)^{\nu_j}} \frac{1}{\mathcal{Z}(\lambda(\mu_j, \nu_j), \nu_j)} \Big) \Big\}. \tag{2.16}$$

The above expression depends on $z_{ij}^{(t)}$, which are referred to as posterior membership probabilities. These arise by noting that $\boldsymbol{Z}_{ij}$ is independent of $Y_{i'}$ for all $i \neq i'$. Since $\mathrm{E}_{\boldsymbol{\Psi}^{(t)}}$ is a linear functional, we may replace $\boldsymbol{Z}_{ij}$ by $\mathrm{E}_{\boldsymbol{\Psi}}[\boldsymbol{Z}_{ij} \mid Y_i = y_i]$, which when

provided the estimate $\boldsymbol{\Psi}^{(t)}$ yields

$$z_{ij}^{(t)} = \frac{\pi_j^{(t)} \left(\frac{\lambda(\mu_j^{(t)}, \nu_j^{(t)})^{y_i}}{(y_i!)^{\nu_j^{(t)}}} \frac{1}{\mathcal{Z}(\lambda(\mu_j^{(t)}, \nu_j^{(t)}), \nu_j^{(t)})} \right)}{\sum_{k=1}^{m} \pi_k^{(t)} \left(\frac{\lambda(\mu_k^{(t)}, \nu_k^{(t)})^{y_i}}{(y_i!)^{\nu_k^{(t)}}} \frac{1}{\mathcal{Z}(\lambda(\mu_k^{(t)}, \nu_k^{(t)}), \nu_k^{(t)})} \right)}. \tag{2.17}$$

M-Step The maximization of $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(t)})$ with respect to $\boldsymbol{\Psi}$ gives the updated estimates $\boldsymbol{\Psi}^{(t+1)}$. First, through the direct use of a Lagrange multiplier, the updated mixing proportions are derived as

$$\frac{\partial Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(t)})}{\partial \pi_j} \stackrel{\text{set}}{=} 0 \Rightarrow \pi_j^{(t+1)} = \frac{\sum_{i=1}^{n} z_{ij}^{(t)}}{n}. \tag{2.18}$$

The updated mean parameter $\mu_j^{(t+1)}$ for component j is then found as a weighted mean of the observations, where the current values of the posterior membership probabilities are the weights:

$$\frac{\partial Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(t)})}{\partial \mu_j} \stackrel{\text{set}}{=} 0 \Rightarrow \mu_j^{(t+1)} = \frac{\sum_{i=1}^{n} z_{ij}^{(t)} y_i}{\sum_{i=1}^{n} z_{ij}^{(t)}}. \tag{2.19}$$

Finally, the updated dispersion parameter $\nu_j^{(t+1)}$ for component j can be obtained by solving

$$\begin{aligned} &\frac{\partial Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(t)})}{\partial \nu_j} \stackrel{\text{set}}{=} 0 \Rightarrow \\ &\frac{\mathrm{E}[Y\log(Y!)] - \mu_j^{(t+1)}\mathrm{E}[\log(Y!)]}{V(Y)} \sum_{i=1}^{n} [z_{ij}^{(t)}(y_i - \mu_j^{(t+1)})] - \sum_{i=1}^{n} [z_{ij}^{(t)} \log(y_i!)] \\ &\quad + \mathrm{E}[\log(Y!)] \sum_{i=1}^{n} z_{ij}^{(t)} = 0. \end{aligned} \tag{2.20}$$

The E-step and M-step are alternated until the stopping criterion $\ell_o(\boldsymbol{\Psi}^{(t+1)}) - \ell_o(\boldsymbol{\Psi}^{(t)}) < \epsilon$, for some small fixed $\epsilon > 0$. Since the MCMP1 distribution is from the exponential family (Huang 2017), the convergence of estimates is always guaranteed

using the EM algorithm (Wu 1983). Also, for completeness, we have supplied the derivations for the mean estimator $\hat{\mu}$ and dispersion estimator $\hat{\nu}$ for the unicomponent MCMP1 distribution in Appendix A.

2.5 Simulation Study

To evaluate the proposed mixtures of MCMP1 model along with our EM algorithm, we generated data from various mixtures of MCMP1 models. In our simulation, different sample sizes and $m \in \{2, 3\}$ components are considered. The results demonstrate the efficacy of the mixture of MCMP1 models under these conditions. The simulated data from two components of MCMP1 mixtures were fit by the MCMP1, negative binomial, and Poisson mixtures for comparison. All numerical work is performed using the R programming language. Flowcharts illustrating the simulation processes used in Sects. 2.5.1 and 2.5.2 are given in Fig. S1 of the Supplementary Information.

2.5.1 Parameter Estimates

We start with a numerical investigation into the biases and root mean squared errors (RMSEs) using our EM algorithm results when fitting various mixtures of MCMP1 models. Case I is comprised of a distribution with two well-separated components with the following parameters: mixing proportions $(\pi, 1-\pi) = (0.3, 0.7)$, means $(\mu_1, \mu_2) = (1, 10)$, and dispersions $(\nu_1, \nu_2) = (0.6, 1.5)$. Case II has two overlapping components with the same parameters as Case I except the means, which are set as $(\mu_1, \mu_2) = (1, 6)$. Case III has a distribution consisting of a well-separated scenario with the following parameters: mixing proportions $(\pi_1, \pi_2, 1-\pi_1-\pi_2) = (0.2, 0.3, 0.5)$, means $(\mu_1, \mu_2, \mu_3) = (1, 10, 25)$, and dispersions $(\nu_1, \nu_2, \nu_3) = (0.6, 1.5, 1.7)$. Comparably, Case IV has the same parameters as Case III except the means, which are set as $(\mu_1, \mu_2, \mu_3) = (1, 6, 25)$. Thus, the three components appear as two overlapping plus one that is more separated. Again, we highlight that a benefit with using the

MCMP1 distribution is that the component mean parameters are at the center of their respective component, and thus can be used to identify the location of each component's distribution. Plots of the mass functions for each of these four cases are given in Fig. S2 of the Supplementary Information.

When assessing the biases and RMSEs, we generate 1000 replications from each of the four MCMP1 mixture models for each of the sample sizes $n \in \{50, 100, 200\}$. One consideration whenever conducting such a simulation study performing optimization via an EM algorithm, especially when estimating finite mixture models, is the choice of initial values. In the present setting, we used a combination of uninformative starting values for the mixing proportions (i.e., set each mixing proportion to $1/m$) and the dispersion parameters (i.e., set each dispersion parameter to 1), but started the algorithm at the component-specific means. Granted the component-specific means are unknown to the user, but since this is a univariate setting, it is expected that the user can posit some reasonable guesses for the component means by either assessing a histogram or proceeding with some other naïve clustering prior to estimating a mixture of MCMP1 model. Regardless of how we proceeded, it always produced the best fitting solution relative to simply generating random starting values for all of the parameters. This is consistent with the focus in this part of the investigation, which is to get the "best" fitting solution so that we can assess the biases and RMSEs.

Tables 2.1 and 2.2 summarize the simulation results for, respectively, the two-component (Cases I and II) and three-component (Cases III and IV) MCMP1 mixture models used in this study. In all cases, we imposed an identifiability constraint where we ordered the means (i.e., $\mu_1 < \mu_2$ in the two-component setting and $\mu_1 < \mu_2 < \mu_3$ in the three-component setting) so as to avoid the *label switching problem* (see Redner and Walker 1984). In both tables, we see that the biases for the mixing proportions and component means across the four cases are all relatively small with a mostly decreasing trend in terms of their absolute values relative to n increasing. However, the

Table 2.1: The average biases and RMSEs from 1000 datasets for various two-component mixtures of MCMP1 models

	Mixing Proportion				Mean				Dispersion			
Case	Parameters	n	Bias	RMSE	Parameters	n	Bias	RMSE	Parameters	n	Bias	RMSE
I[a]	$\pi_1 = 0.3$	50	0.0008	0.0669	$\mu_1 = 1$	50	-0.0126	0.3559	$\nu_1 = 0.6$	50	0.9172	2.2789
		100	0.0002	0.0468		100	-0.0127	0.2473		100	0.3260	0.7862
		200	-0.0011	0.0332		200	-0.0056	0.1636		200	0.1697	0.4319
	$\pi_2 = 0.7$	50	-0.0008	0.0669	$\mu_2 = 10$	50	0.0004	0.4642	$\nu_2 = 1.5$	50	0.1699	0.5186
		100	-0.0002	0.0468		100	-0.0265	0.3131		100	0.0809	0.3382
		200	0.0011	0.0332		200	0.0035	0.2280		200	0.0327	0.2233
II[b]	$\pi_1 = 0.3$	50	0.0002	0.0954	$\mu_1 = 1$	50	-0.0509	0.5273	$\nu_1 = 0.6$	50	1.6722	3.4699
		100	0.0009	0.0695		100	-0.0240	0.4066		100	0.8495	2.2372
		200	0.0000	0.0453		200	-0.0023	0.2600		200	0.3495	1.1962
	$\pi_2 = 0.7$	50	-0.0001	0.0954	$\mu_2 = 6$	50	-0.0041	0.4788	$\nu_2 = 1.5$	50	0.2670	0.7316
		100	-0.0009	0.0695		100	0.0030	0.3355		100	0.1471	0.4836
		200	0.0000	0.0453		200	0.0020	0.2249		200	0.0738	0.2742

[a]Case I exemplifies two separate components of MCMP1s.
[b]Case II exemplifies two overlapping components of MCMP1s.

Table 2.2: The average biases and RMSEs from 1000 datasets for various three-component mixtures of MCMP1 models

	Mixing Proportion				Mean				Dispersion			
Case	Parameters	n	Bias	RMSE	Parameters	n	Bias	RMSE	Parameters	n	Bias	RMSE
III[a]	$\pi_1 = 0.2$	50	0.0000	0.0590	$\mu_1 = 1$	50	-0.0084	0.4474	$\nu_1 = 0.6$	50	1.5907	3.3510
		100	-0.0009	0.0415		100	-0.0130	0.3048		100	0.5541	1.3096
		200	-0.0011	0.0280		200	-0.0072	0.2177		200	0.2533	0.5891
	$\pi_2 = 0.3$	50	0.0040	0.0692	$\mu_2 = 10$	50	-0.0394	0.8530	$\nu_2 = 1.5$	50	0.3174	0.8001
		100	0.0009	0.0496		100	-0.0243	0.5710		100	0.1853	0.6257
		200	0.0018	0.0351		200	0.0052	0.3749		200	0.0983	0.4512
	$\pi_3 = 0.5$	50	-0.0040	0.0727	$\mu_3 = 25$	50	-0.0074	0.8135	$\nu_3 = 1.7$	50	0.0725	0.3636
		100	-0.0001	0.0522		100	-0.0112	0.5692		100	0.0476	0.2991
		200	-0.0007	0.0369		200	0.0154	0.4071		200	0.0456	0.2504
IV[b]	$\pi_1 = 0.2$	50	-0.0030	0.0731	$\mu_1 = 1$	50	-0.0413	0.5932	$\nu_1 = 0.6$	50	2.1516	4.0675
		100	-0.0038	0.0540		100	-0.0563	0.4605		100	1.3857	3.0469
		200	-0.0026	0.0401		200	-0.0388	0.3230		200	0.6082	1.7413
	$\pi_2 = 0.3$	50	0.0021	0.0772	$\mu_2 = 6$	50	0.0080	0.7359	$\nu_2 = 1.5$	50	0.5294	1.0987
		100	-0.0000	0.0594		100	-0.0168	0.5610		100	0.3023	0.8541
		200	0.0034	0.0433		200	-0.0247	0.3713		200	0.1074	0.5247
	$\pi_3 = 0.5$	50	0.0008	0.0720	$\mu_3 = 25$	50	-0.0473	0.7767	$\nu_2 = 1.7$	50	0.0624	0.3491
		100	0.0037	0.0508		100	0.0215	0.5453		100	0.0532	0.2867
		200	-0.0008	0.0354		200	0.0215	0.3928		200	0.0398	0.2357

[a]Case III exemplifies three separate components of MCMP1s.

[b]Case IV exemplifies three components as two overlapping plus one more-separated of MCMP1s.

bias values for the dispersion parameters are slightly larger, but clearly demonstrate an overall decreasing pattern as n increases. In all four cases, the RMSEs decrease as n increases for the mixing proportions, means, and dispersions. However, the RMSEs for the mixing proportions and means are smaller relative to the dispersions. The bias and RMSE results noted about the dispersion parameters are likely attributable to the larger or smaller variance that is expected when a component is, respectively, over-dispersed or under-dispersed, such that the MCMP1 affords us with the flexibility of reasonably characterizing the latter. Overall, these results for the component-specific dispersion parameters indicate that the sampling variability has a noticeable impact on the accuracy of estimating this parameter. This is further highlighted by the results for the component with a larger mixing proportion, which yields smaller RMSEs for the corresponding dispersion parameter. For both the two-component (Table 2.1) and three-component (Table 2.2) mixtures of MCMP1s, the RMSEs in the overlapping cases are higher than that in the well-separated cases. Overall, these results indicate that our EM algorithm is producing consistent estimates for the MCMP1 mixture models considered in this study.

2.5.2 Model Selection

We now turn to a simulation study to assess how well other competitor models fit data that were, in fact, generated from a mixture of MCMP1s. Specifically, we generate data from two-component mixtures of MCMP1s with varying degrees of dispersion, and then compare the corresponding fits to those obtained under two-component mixtures of Poissons and two-component mixtures of negative binomials. Estimation of mixtures of MCMPs follows what we presented in Sect. 2.4. The `glm.nb()` function in the `MASS` package (Venables and Ripley 2002) is used to construct an EM algorithm for estimating mixtures of negative binomials. We use the `flexmix()` function in the `flexmix` package (Leisch 2004) for estimating mixtures of Poissons. We found that

the solutions obtained for each model fit did not heavily depend on the choice of initial values since the components of the data-generating models (to be defined shortly) are well-separated. Thus, we just use a simple k-means method in R to roughly group the data for having the initial values to start the EM algorithm in the mixtures of CMP and mixture of negative binomial models.

After fitting the candidate mixture models, we proceed to use the Bayesian information criterion (BIC; Schwarz 1978) to select the best fitting model. Recall that the BIC values are calculated using the formula $-2\ell_o^{(\infty)} + d\log(n)$, where $\ell_o^{(\infty)}$ is the converged (observed) loglikelihood from the corresponding optimization routine, d is the number of parameters in the model, and n is the sample size. Thus, when comparing across a set of candidate models, a smaller BIC value indicates a better-fitting model.

The use of information criteria for selecting the number of components in finite mixtures is well-studied; see, for example, Chapter 6 of McLachlan and Peel (2000). Akaike's information criterion (AIC; Akaike 1973) is also frequently employed for the broader purpose of model selection, which includes choosing among candidate count models. However, there is a breakdown in regularity conditions for the asymptotic expansions of the penalty terms for many of these information criteria. The implications of this on selecting the number of components in a mixture model is that AIC is order-inconsistent and tends to overestimate the correct number of components (Celeux and Soromenho 1996). BIC, however, has empirically been shown to perform well, including that it does not underestimate the true number of components, asymptotically (Leroux 1992). Moreover, the use of BIC is generally supported in the mixture modeling literature (Fraley and Raftery 1998). Thus, we proceed with interpreting the BIC results for our simulation work and data analysis, but include in the Supplementary Information summaries of the loglikelihood and AIC values for completeness.

We generate data from two-component mixture of MCMP1 models having mean parameters $(\mu_1, \mu_2) = (1, 15)$ and mixing proportions $(\pi, 1-\pi) = (0.3, 0.7)$. Such components are well-separated for most values of dispersion, which helps us to avoid any potential computational difficulties when fitting the different models. For this simulation, we consider varying amounts of dispersion, (ν_1, ν_2), as well as the two sample sizes $n \in \{100, 200\}$. The values used for the dispersion parameters in the two-component mixture of MCMP1 models, as well as the comparative boxplots of the corresponding BIC results, are as follows:

- $(\nu_1, \nu_2) = (0.4, 0.5)$; Fig. 2.1a ($n = 100$) and e ($n = 200$);
- $(\nu_1, \nu_2) = (1.4, 1.5)$; Fig. 2.1b ($n = 100$) and g ($n = 200$);
- $(\nu_1, \nu_2) = (2.4, 2.5)$; Fig. 2.1c ($n = 100$) and g ($n = 200$);
- $(\nu_1, \nu_2) = (3.4, 3.5)$; Fig. 2.1d ($n = 100$) and h ($n = 200$).

As in our previous simulation study, we again generate 1000 such datasets for each simulation condition. Histograms of a simulated dataset from each of the above settings, as well as the corresponding fits for each of the three mixture models considered for this study, are given in Fig. S3 of the Supplementary Information.

Figure 2.1a and e concern the settings where both components have over-dispersion, as indicated by their dispersion parameters $(\nu_1, \nu_2) = (0.4, 0.5)$ both being less than one. In these settings, mixtures of MCMP1s and mixtures of negative binomials both provide better fits than mixtures of Poissons. This is noted by both boxplots being slightly lower relative to those for the mixture of Poissons setting. However, the mixtures of MCMP1s and mixtures of negative binomials have similar boxplots, which is typical as the negative binomial and CMP distributions tend to both fit over-dispersed data similarly.

In the MCMP1, a larger dispersion parameter corresponds to an increased amount of under-dispersion. We show the results of increased dispersion parameters (greater

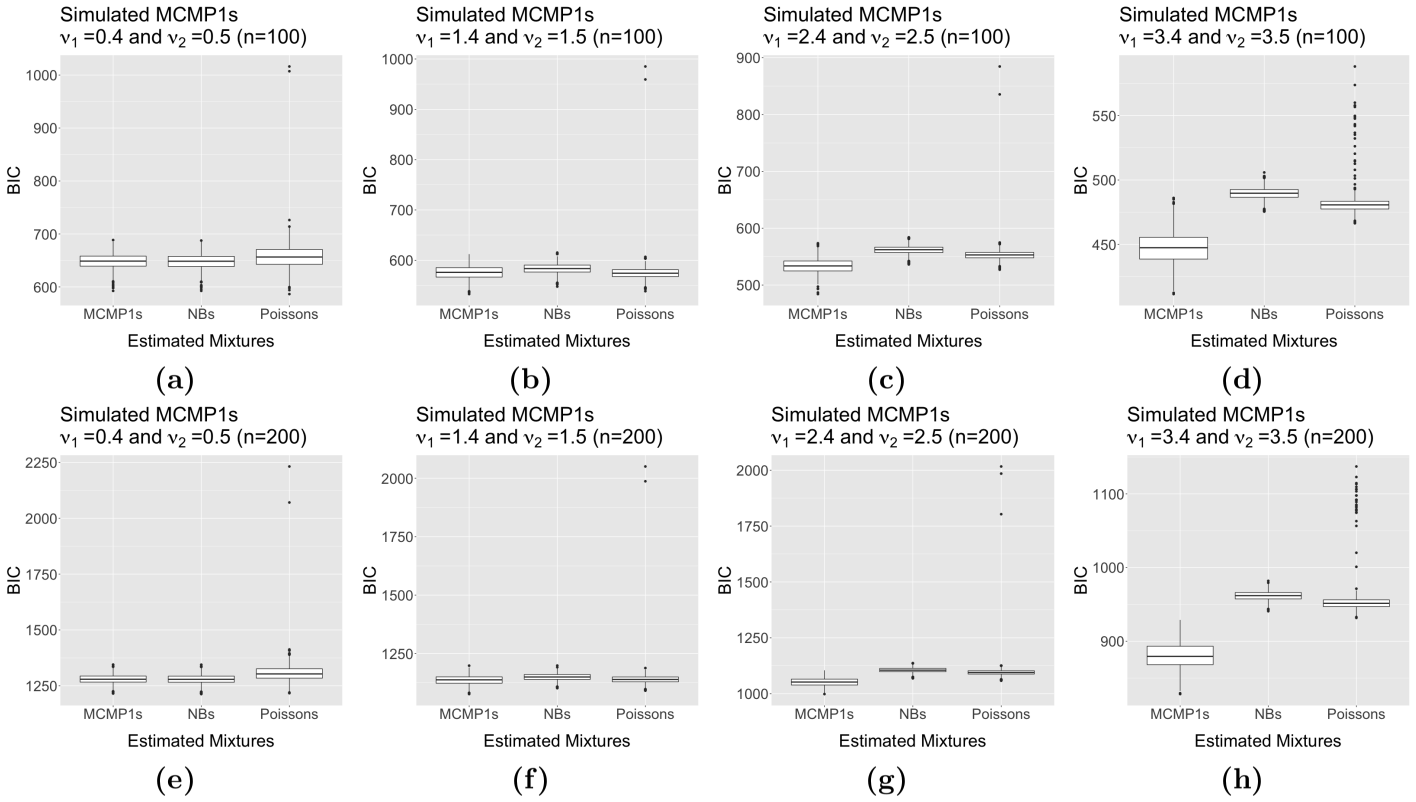


Figure 2.1: Comparative boxplots of BIC values associated with two-component mixtures of MCMP1s, two-component mixtures of negative binomials (NBs), and two-component mixtures of Poissons fitted to the simulated data. Each boxplot summarizes the BIC values for 1000 datasets generated from two-component mixtures of MCMP1s. Plots (a)–(d) summarize datasets of size $n = 100$ and plots (e)–(h) summarize datasets of size $n = 200$. The following dispersion parameters are used in the simulation: $(0.4, 0.5)$ for plots (a) and (e), $(1.4, 1.5)$ for plots (b) and (f), $(2.4, 2.5)$ for plots (c) and (g), and $(3.4, 3.5)$ for plots (d) and (h).

amounts of under-dispersion) with $(1.4, 1.5)$ in Fig. 2.1b and f, $(2.4, 2.5)$ in Fig. 2.1c and g, and $(3.4, 3.5)$ in Fig. 2.1d and h. Specifically, it becomes more apparent that mixtures of MCMP1s have the advantage over both negative binomial and Poisson mixtures as the boxplots of their BIC values are the lowest among the three estimated mixture models. For the three under-dispersed cases, the mixture of negative binomials models is not even better than the mixture of Poissons model, which indicates that a mixture of negative binomials is not competitive at fitting mixtures with components that consist of under-dispersed data. The mixture of Poissons model has right-skewed BIC values, indicating that data generated from a mixture with components having dispersion does not fit well with a mixture of Poissons. These results are consistent across the sample sizes $n = 100$ (Fig. 2.1a–d and $n = 200$ (Fig. 2.1e–h).

For completeness, Figs. S4 and S5 in the Supplementary Information show the analogous set of comparative boxplots for, respectively, the loglikelihood values and AIC values. Both sets of figures demonstrate the same type of behavior as what we observed with the BIC values in Fig. 2.1.

2.6 Application: Dog Mortality Data

Lewis et al. (2018) presented statistical summaries about the age at death for $n = 5663$ dogs across 179 breeds from a mortality survey administered by The Kennel Club in the United Kingdom. The ages at death are reported in years as integers, ranging from a minimum of 0 years to a maximum 26 years. Thus, it is appropriate to investigate modeling of these data using discrete distributions. Additional covariate information for each dog (e.g., breed and cause of death) is not available in the reduced version of the dataset provided, so proceeding with a univariate analysis is necessary. Moreover, the histogram of the dog death ages in years (Fig. 2.2) clearly demonstrates multimodality, which would also indicate that a mixture of discrete distributions is appropriate. A mixture is also practically appropriate because missing

covariate information (like the breed of dog) would likely inform some of the different subpopulations of ages at time of death. For example, larger dog breeds like the Irish Wolfhound are known to have very short lifespans (about 7–10 years), whereas smaller breeds like the Chihuahua are known to have relatively longer lifespans (often exceeding 12 years).

For the simulation study in Sect. 2.5, we took a mostly informative strategy for specifying the initial values of our EM algorithm for estimating the mixture of MCMP1 model. When initializing our EM algorithm at or near the true component means, excellent results were obtained in the simulation study. Those component means are also where the different peaks occur, assuming that the components are not heavily overlapping. In a real data analysis, it is most advantageous to choose the initial values for the means in our EM algorithm by postulating the location of the peaks, which will be a strong indicator of the underlying component means. Notwithstanding, we assessed the estimated model results using different initial values on the dog mortality data in order to identify the best fit.

As in the simulation study, we use the `flexmix()` function for estimating mixtures of Poissons and developed an EM algorithm that uses the `glm.nb()` function for estimating mixtures of negative binomials. The `flexmix()` function has the capability of using random starting values and we can employ a similar random starting value strategy with our EM algorithm. We fit the dog death data by performing 100 random initializations of these algorithms for both mixture models with $m \in \{2, 3, 4\}$ components. The results for a given model were consistently comparable, with differences appearing mostly on the order of 10^{-3}. Regardless, the one with the largest loglikelihood, which corresponds to the smallest BIC, among the 100 replications is chosen to represent the case in Table 2.3. The results for the non-mixture setting (i.e., $m = 1$) are also reported.

For using our EM algorithm to estimate the parameters of a mixture of MCMP1

model (again, for $m \in \{1, 2, 3, 4\}$), we identified a grid of candidate mean values to use as our initial values. Based on the range of the dog death ages, an equally-spaced sequence consisting of seven elements is created as $\Omega = \{0, 4.3, 8.7, 13, 17.3, 21.7, 26\}$. One, two, three, or four unique values are chosen from Ω as the initial component means to start the EM algorithm for the respective mixture of MCMP1 model having m components. Overall, there are 7, 21, 35, and 35 distinct combinations of starting values for the means for, respectively the unicomponent, two-component, three-component, and four-component models. Note that starting values are necessary for the optimization algorithm for the unicomponent model, where the formulas used for performing maximum likelihood estimation of that model are highlighted in Appendix A. The stopping criterion (based on the difference between the observed loglikelihoods of successive EM iterations as discussed at the end of Sect. 2.4) is set at $\epsilon = 10^{-3}$. The result with the largest loglikelihood is chosen among the fits to calculate the BIC for each of the four mixtures of MCMP1 models summarized in Table 2.3.

When comparing the BIC values of the estimated mixtures having either Poisson or negative binomial components, the results in Table 2.3 show that the two-component models are the best. However, when comparing those along with all of the mixtures of MCMP1s, we see that the three-component mixture of MCMP1 has the smallest BIC, which is indicated in boldface. In fact, for $m \in \{2, 3, 4\}$ components,

Table 2.3: BIC values for mixtures of Poissons, mixtures of negative binomials (NBs), and mixtures of MCMP1s fitted to the dog death data

	BIC		
m	Poissons	NBs	MCMP1s
1	33994.3019	32844.9115	32403.3754
2	31432.6595	31435.9771	31284.1782
3	31449.9429	31461.8959	**31194.9380**
4	31467.2277	31483.4729	31205.8337

the Poisson mixtures are better than the respective negative binomial mixtures. The negative binomial does a good job characterizing over-dispersion in data, so having a mixture of Poissons perform better than a mixture of negative binomials is indicative that the components are either (mostly) exhibiting equi-dispersion or, possibly, under-dispersion. In this case, the fact that the mixture of MCMP1s is the best according to the BIC values is indicative that the data is demonstrating the latter.

We next visualize the results for the three-component mixture fits. Figure 2.2 is a histogram of the dog mortality data (in years at time of death), with the estimated three-component mixture of MCMP1s (solid line), negative binomials (dotted line), and Poissons (dashed line). Note that while these are discrete distributions and, thus, supported on $\mathbb{N}$, we have connected the estimated mass values so as to more easily discern the shapes of the different mixture model fits. The negative binomial and Poisson mixtures are practically overlaid on top of one another. Their second component (where the peak appears at an age of 12 years) appears to exhibit some lack of fit as the shape is shifted to the left of the raw data in this part of the histogram. In contrast, the MCMP1 mixture appears to do much better at capturing the shape of the histogram, including the presence of a third component starting to emerge

Figure 2.2: Histogram of the observed dog death ages of Lewis et al. (2018) and the fits for the three-component mixtures of MCMP1s (solid line), negative binomials (dotted line), and Poissons (dashed line)

around a value of 8 years. This is further illustrated by the parameter estimates for these three models, which are given in Table 2.4. For the estimated Poisson and negative binomial mixture models, the second and third components effectively have the same means, which underscores the inability of these two models to be able to clearly distinguish the presence of a third component. Moreover, the mixture of MCMP1 model picks up three components, each with quite different dispersions: 1.8, 0.7, and 2.7. Meanwhile, the negative binomial mixture only discerns an over-dispersed component with a dispersion estimate of 5.9, but the other two components are effectively equi-dispersed due to their extremely large dispersion estimates. This results in the variance being approximately equal to the mean in each of the two components.

Table 2.4: The estimated mixing proportions ($\widehat{\pi}_j$), means ($\widehat{\mu}_j$), and dispersions ($\widehat{\alpha}_j, \widehat{\nu}_j$), $j = 1, 2, 3$, for the estimated three-component mixtures of MCMP1s, negative binomials (NBs), and Poissons

	$\widehat{\pi}_j$			$\widehat{\mu}_j$			$\widehat{\alpha}_j$ (NB), $\widehat{\nu}_j$ (MCMP1)		
Component j	1	2	3	1	2	3	1	2	3
Poisson	0.13	0.43	0.44	2.54	10.73	10.73	—	—	—
NB	0.14	0.43	0.43	2.98	10.80	10.82	5.9	156338	157683
MCMP1	0.06	0.48	0.46	1.42	8.28	12.34	1.8	0.7	2.7

We next provide more specific interpretations as they pertain to the estimated mixture of MCMP1s model. The corresponding parameter estimates are given (again) in Table 2.5, but this time also with estimated standard errors. From these results, we can state that approximately 6.5% of the dogs in this survey belong to the first component, which has the smallest mean estimate of 1.42 years at the time of death. This component likely consists of dogs coming from a wide array of breeds, but where they died as puppies or in young adulthood due to various life-ending congenital abnormalities or possible accidents. Next, we can state that approximately 47.83% of the dogs in this survey belong to the second component, which has a mean estimate

of 8.28 years. This component is likely comprised of mostly larger-breed dogs, which typically have shorter lifespans. Finally, we can state that 45.67% of the dogs in this survey belong to the third component, which has the largest mean estimate of 12.34 years at the time of death. This component is likely comprised of mostly smaller-breed dogs, which often have longer lifespans. As briefly mentioned earlier, these components show distinct dispersion behavior, with estimates 1.8, 0.7, and 2.7, respectively. In the MCMP1 distribution, this means that the first and third components exhibit under-dispersion (as they are greater than 1), while the second component exhibits over-dispersion (as it is less than 1).

In Table 2.5, we also provide estimated standard errors for each of the parameter estimates. We estimated the standard errors using both the traditional non-parametric and parametric bootstrap. For both approaches, we drew $B = 1000$ bootstrap samples. For each bootstrap sample, we set the starting values of our EM algorithm to the parameter estimates that are reported in Table 2.5. Both bootstraps yield similar standard errors for a given parameter estimate, which are much smaller in magnitude comparing to their corresponding estimates. Of particular note is that the intervals of $\widehat{\mu}_j \pm (3 \times \widehat{\mathrm{SE}}(\widehat{\mu}_j))$, $j = 1, 2, 3$, do not overlap, thus indicating that the means of the components are (significantly) different.

Table 2.5: The parameter estimates for the three-component mixture of MCMP1s as also reported in Table 2.4, as well as their estimated standard errors ($\widehat{\mathrm{SE}}$s) in parentheses, which are calculated using a parametric (left of the slash) and non-parametric (right of the slash) bootstrap

Component j	$\widehat{\pi}_j \left(\widehat{\mathrm{SE}}(\widehat{\pi}_j)\right)$	$\widehat{\mu}_j \left(\widehat{\mathrm{SE}}(\widehat{\mu}_j)\right)$	$\widehat{\nu}_j \left(\widehat{\mathrm{SE}}(\widehat{\nu}_j)\right)$
1	0.0650 (0.0093/0.0097)	1.4202 (0.2147/0.2201)	1.8 (0.3399/0.2456)
2	0.4783 (0.0194/0.0218)	8.2765 (0.2605/0.2800)	0.7 (0.0467/0.0512)
3	0.4567 (0.0231/0.0266)	12.3432 (0.0791/0.0651)	2.7 (0.1222/0.0150)

As noted earlier, this survey data consists of 179 unique breeds as recognized by The Kennel Club in the United Kingdom. Given that domestic dogs are quite

diverse with large variation among breeds, the MCMP1 mixture provides a potential tool to classify the dogs according to their lifespans, which in this dataset consists of those with a short lifespan (1.42 years), medium lifespan (8.28 years), or long lifespan (12.34 years). Beyond considering all breeds as a whole, the statistics we cited for the components provide a more nuanced view about the lifespan of dogs that are registered with The Kennel Club.

2.7 Discussion

The CMP distribution is a flexible generalization of the Poisson distribution, and has seen a resurgence in popularity over about the past 20 years [see, for example, the references within the text by Sellers (2023)]. One possible impediment to the wider use of the classic CMP model in practice is the inability to directly model the mean of the distribution. Huang (2017) remedied this issue by introducing the MCMP1 distribution, which was used in our work.

One setting where there is scant treatment of CMP distributions is in the context of mixture modeling. Specifically, such a model can allow discrete data for a particular application to be comprised of subpopulations, each possibly having varying degrees of dispersion. As we noted, the only relevant mixture model to our knowledge appears in Sur et al. (2015). However, that work only addresses a two-component mixture of truncated CMPs for the purpose of characterizing the datasets presented in that work, all of which are truncated below at $t = 1$. Moreover, those authors did not present estimated standard errors for their estimates, which we addressed through bootstrapping. Overall, our work fills the gap of having a more general finite mixture model with components that are (mean-parameterized) CMP distributions.

We also developed an EM algorithm for estimating our finite mixture of MCMP1 models. This algorithm was employed in the extensive numerical study of Sect. 2.5. These results not only showed the excellent performance of our algorithm for

estimating our proposed mixture model, but we also demonstrated its relative competitiveness, and in certain cases superiority, when modeling data that arise as a mixture of discrete distributions with varying degrees of dispersion. This was further demonstrated in the analysis of the dog mortality data of Sect. 2.6, where we identified three possible subpopulations in the dataset, and also showed that the MCMP1 mixture model is better (in terms of its BIC value) than the Poisson and negative binomial mixture models. Thus, our mixture of MCMP1 model contributes a meaningful extension to the flexible class of CMP-based models.

One limitation with what we presented is that our model is only for the univariate setting. We are currently developing the regression extension of our model to incorporate covariates. Specifically, using a GLM framework, we can model each of the component parameters, μ_j, ν_j, and π_j, as a function of covariates. This could be done by using a log link for μ_j and ν_j, and a logit link for π_j. Of course, one big issue here will be to investigate identifiability of such a generalized mixture of MCMP1 regressions model. We expect to have results and computational routines for this research to present in the near future.

2.8 Appendix A: ML Estimators of Mean-Parameterized CMP Distribution

Following the work of Huang (2017), this appendix gives the derivations of the maximum likelihood estimators (MLEs) for μ and ν in the MCMP1 distribution (2.4).

We first note that the normalizing constant $\mathcal{Z}(\lambda, \nu)$ in the CMP density function (2.4) is a function of λ and ν, so the partial derivatives of $\mathcal{Z}(\lambda, \nu)$ are required in the derivation of the MLEs for μ and ν in the MCMP1 distribution. Taking the partial

derivative of $\mathcal{Z}(\lambda,\nu)$ with respect to λ, we have

$$\begin{aligned}\frac{\partial \mathcal{Z}(\lambda,\nu)}{\partial \lambda} &= \sum_{y=0}^{\infty} \frac{y\lambda^{(y-1)}}{(y!)^{\nu}} \\ &= \frac{1}{\lambda}\sum_{y=0}^{\infty} \frac{y\lambda^{y}}{(y!)^{\nu}} \\ &= \frac{\mu\mathcal{Z}(\lambda,\nu)}{\lambda}.\end{aligned}$$

Then, taking the partial derivative of $\mathcal{Z}(\lambda,\nu)$ with respect to ν, we have

$$\frac{\partial \mathcal{Z}(\lambda,\nu)}{\partial \nu} = -\sum_{y=0}^{\infty} \frac{\lambda^{y}}{(y!)^{\nu}}\log(y!).$$

Accordingly,

$$\begin{aligned}-\frac{1}{\mathcal{Z}(\lambda,\nu)}\frac{\partial \mathcal{Z}(\lambda,\nu)}{\partial \nu} &= \sum_{y=0}^{\infty} \frac{1}{\mathcal{Z}(\lambda,\nu)}\frac{\lambda^{y}}{(y!)^{\nu}}\log(y!) \\ &= \mathrm{E}[\log(Y!)].\end{aligned}$$

Now, let $Y_1,\ldots,Y_n$ be *iid* MCMP1 random variables with corresponding realizations $y_1,\ldots,y_n$. The loglikelihood function is then given by

$$\ell(\mu,\nu) \equiv \ell = \sum_{i=1}^{n}\Big\{y_i\log\lambda(\mu,\nu) - \nu\log(y_i!) - \log\mathcal{Z}(\lambda(\mu,\nu),\nu)\Big\}. \tag{A1}$$

When reparameterizing the CMP distribution using its mean μ and dispersion ν, the original rate parameter λ in Eq. (2.5) is considered as a function of μ and ν. The following partial derivatives are required in the derivation of the MLEs for μ and ν. First, taking the partial derivative on both sides of Eq. (2.5) with respect to μ, we

have

$$
\begin{aligned}
0 &= \sum_{y=0}^{\infty} \Big\{ -\frac{\lambda(\mu,\nu)^y}{(y!)^\nu} + (y-\mu)\frac{y\lambda(\mu,\nu)^{(y-1)}\frac{\partial\lambda(\mu,\nu)}{\partial\mu}}{(y!)^\nu} \Big\} \\
&= -\sum_{y=0}^{\infty}\frac{\lambda(\mu,\nu)^y}{(y!)^\nu} + \frac{\partial\lambda(\mu,\nu)}{\partial\mu}\sum_{y=0}^{\infty}(y-\mu)y\frac{\lambda(\mu,\nu)^{(y-1)}}{(y!)^\nu} \\
&= -\mathcal{Z}(\lambda(\mu,\nu),\nu) + \frac{1}{\lambda(\mu,\nu)}\frac{\partial\lambda(\mu,\nu)}{\partial\mu}\sum_{y=0}^{\infty}(y-\mu)y\frac{\lambda(\mu,\nu)^y}{(y!)^\nu} \\
&= -\mathcal{Z}(\lambda(\mu,\nu),\nu) + \frac{1}{\lambda(\mu,\nu)}\frac{\partial\lambda(\mu,\nu)}{\partial\mu}\mathcal{Z}(\lambda(\mu,\nu),\nu)\sum_{y=0}^{\infty}(y-\mu)y\frac{\lambda(\mu,\nu)^y}{(y!)^\nu}\frac{1}{\mathcal{Z}(\lambda(\mu,\nu),\nu)} \\
&= -\mathcal{Z}(\lambda(\mu,\nu),\nu) + \frac{1}{\lambda(\mu,\nu)}\frac{\partial\lambda(\mu,\nu)}{\partial\mu}\mathcal{Z}(\lambda(\mu,\nu),\nu)\mathrm{E}[(Y-\mu)Y] \\
&= -\mathcal{Z}(\lambda(\mu,\nu),\nu) + \frac{1}{\lambda(\mu,\nu)}\frac{\partial\lambda(\mu,\nu)}{\partial\mu}\mathcal{Z}(\lambda(\mu,\nu),\nu)V(Y) \\
&\Rightarrow \frac{\partial\lambda(\mu,\nu)}{\partial\mu} = \frac{\lambda(\mu,\nu)}{V(Y)},
\end{aligned}
$$

where $V(Y) = \mathrm{E}[(Y-\mu)Y]$ is the variance of the CMP distribution. Second, taking the partial derivative on both sides of Eq. (2.5) with respect to ν, we have

$$
\begin{aligned}
0 &= \sum_{y=0}^{\infty}(y-\mu)\frac{y\lambda(\mu,\nu)^{(y-1)}\frac{\partial\lambda(\mu,\nu)}{\partial\nu}(y!)^\nu - \lambda(\mu,\nu)^y(y!)^\nu\log(y!)}{(y!)^{2\nu}} \\
&= \frac{1}{\lambda(\mu,\nu)}\frac{\partial\lambda(\mu,\nu)}{\partial\nu}\sum_{y=0}^{\infty}(y-\mu)y\frac{\lambda(\mu,\nu)^y}{(y!)^\nu} - \sum_{y=0}^{\infty}(y-\mu)\log(y!)\frac{\lambda(\mu,\nu)^y}{(y!)^\nu} \\
&= \frac{1}{\lambda(\mu,\nu)}\frac{\partial\lambda(\mu,\nu)}{\partial\nu}\mathcal{Z}(\lambda(\mu,\nu),\nu)\mathrm{E}[(Y-\mu)Y] - \mathcal{Z}(\lambda(\mu,\nu),\nu)\mathrm{E}[(Y-\mu)\log(Y!)] \\
&= \frac{1}{\lambda(\mu,\nu)}\frac{\partial\lambda(\mu,\nu)}{\partial\nu}\mathcal{Z}(\lambda(\mu,\nu),\nu)V(Y) - \mathcal{Z}(\lambda(\mu,\nu),\nu)\mathrm{E}[(Y-\mu)\log(Y!)] \\
&\Rightarrow \frac{\partial\lambda(\mu,\nu)}{\partial\nu} = \frac{\lambda(\mu,\nu)}{V(Y)}\mathrm{E}[(Y-\mu)\log(Y!)].
\end{aligned}
$$

To find the MLE for μ, first take the partial derivative of (A1) with respect to μ

and set it equal to zero:

$$\begin{aligned}
\frac{\partial \ell}{\partial \mu} &= \frac{\partial \ell}{\partial \lambda(\mu,\nu)} \frac{\partial \lambda(\mu,\nu)}{\partial \mu} \\
&= \sum_{i=1}^{n} \left\{ \frac{y_i}{\lambda(\mu,\nu)} - \frac{1}{\mathcal{Z}(\lambda(\mu,\nu),\nu)} \frac{\partial \mathcal{Z}(\lambda(\mu,\nu),\nu)}{\partial \lambda(\mu,\nu)} \right\} \frac{\partial \lambda(\mu,\nu)}{\partial \mu} \\
&= \sum_{i=1}^{n} \left\{ \frac{y_i}{\lambda(\mu,\nu)} - \frac{1}{\mathcal{Z}(\lambda(\mu,\nu),\nu)} \frac{\mu \mathcal{Z}(\lambda(\mu,\nu),\nu)}{\lambda(\mu,\nu)} \right\} \frac{\lambda(\mu,\nu)}{V(Y)} \\
&= \frac{\sum_{i=1}^{n} y_i - n\mu}{V(Y)} \\
&\stackrel{\text{set}}{=} 0.
\end{aligned}$$

Then, it immediately follows that the MLE $\hat{\mu}$ is solved as the mean of the observed sample; i.e.,

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y_i. \tag{A2}$$

To find the MLE for ν, we again take the first partial derivative of (A1), but with respect to ν and set it equal to zero:

$$\begin{aligned}
\frac{\partial \ell}{\partial \nu} &= \sum_{i=1}^{n} \left\{ \frac{y_i}{\lambda(\mu,\nu)} \frac{\partial \lambda(\mu,\nu)}{\partial \nu} - \log(y_i!) - \frac{1}{\mathcal{Z}(\lambda(\mu,\nu),\nu)} \frac{\partial \mathcal{Z}(\lambda(\mu,\nu),\nu)}{\partial \nu} \right\} \\
&= \sum_{i=1}^{n} \left\{ \frac{(y_i - \mu)}{\lambda(\mu,\nu)} \frac{\lambda(\mu,\nu)}{V(Y)} \mathrm{E}[(Y-\mu)\log(Y!)] - \log(y_i!) + \mathrm{E}[\log(Y!)] \right\} \\
&= \sum_{i=1}^{n} \left\{ \frac{(y_i - \mu)}{V(Y)} \mathrm{E}[(Y-\mu)\log(Y!)] - \log(y_i!) + \mathrm{E}[\log(Y!)] \right\} \\
&\stackrel{\text{set}}{=} 0.
\end{aligned}$$

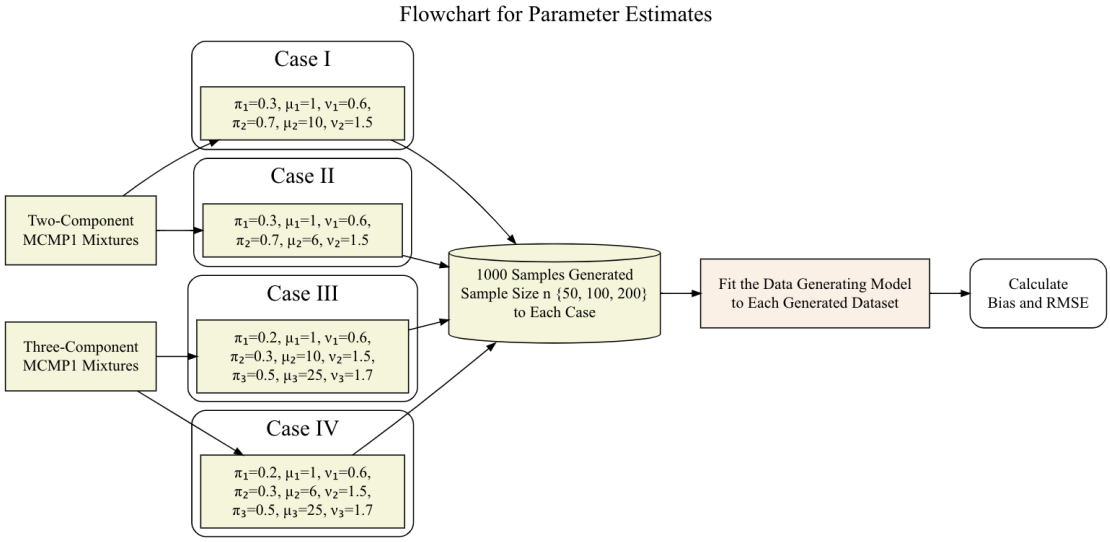
Then, the MLE $\hat{\nu}$ is found as the solution that satisfies

$$\frac{\mathrm{E}[Y\log(Y!)] - \mu \mathrm{E}[\log(Y!)]}{V(Y)} (\sum_{i=1}^{n} y_i - n\mu) - \sum_{i=1}^{n} \log(y_i!) + n\mathrm{E}[\log(Y!)] = 0, \tag{A3}$$

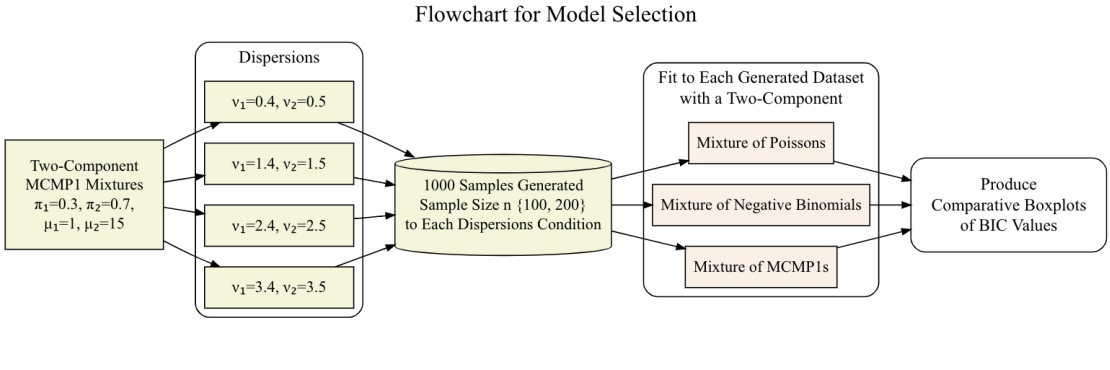
which requires use of a numerical routine.

2.9 Appendix B: Additional Figures and Numerical Results

In this supplementary file, we provide flowcharts and additional results for the simulation study conducted in Section 2.5 of the main text as well as additional results for the analysis of the dog mortality data in Section 2.6.



(a)



(b)

Figure S1: Flowcharts for (a) the numerical study about parameter estimates in Section 2.5.1 and (b) the model selection study in Section 2.5.2

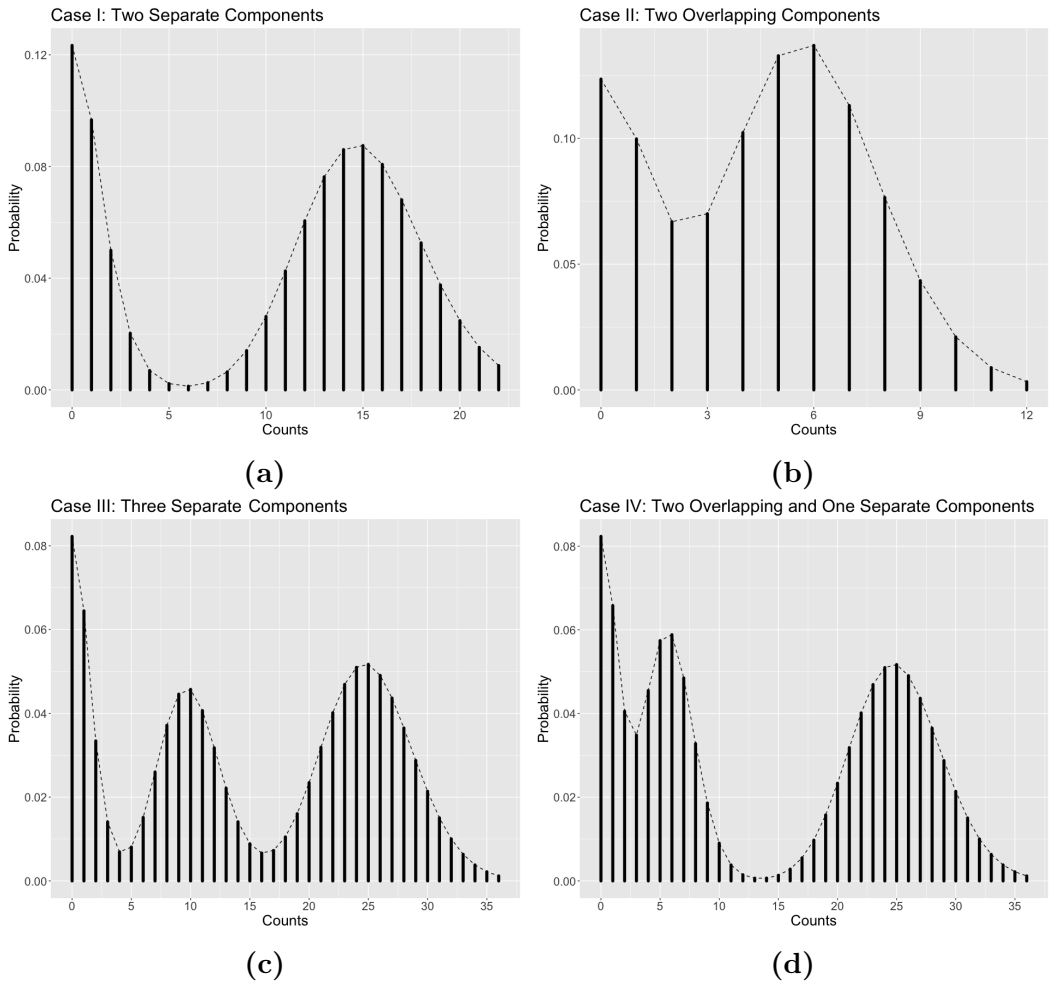


Figure S2: Mass functions for the mixtures of MCMP1 models for (a) Case I, (b) Case II, (c) Case III, and (d) Case IV, as defined in the main text

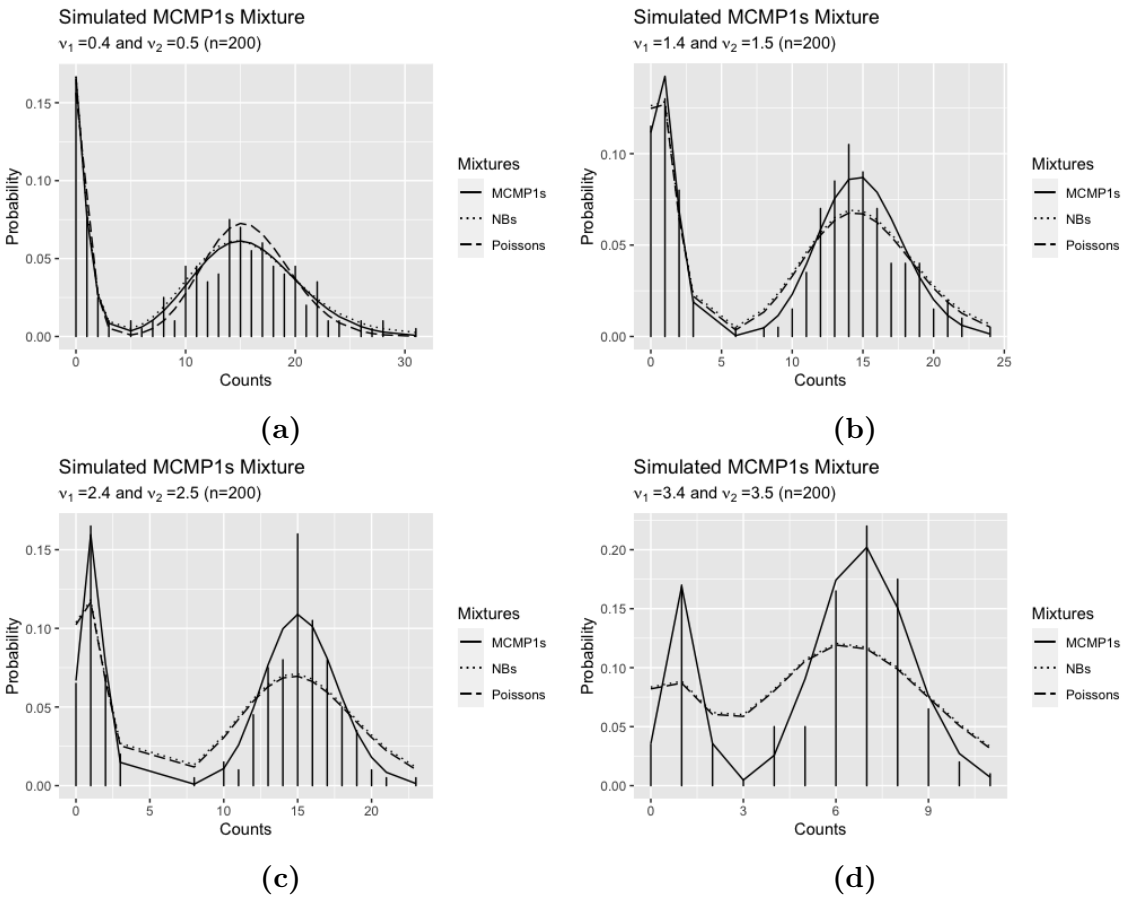


Figure S3: Simulated datasets from each two-component mixture of MCMP1 model used in the model selection study in the main text. The fitted two-component mixture of MCMP1s (solid line), two-component mixture of negative binomials (dotted line), and two-component mixture of Poissons (dashed line) are each overlaid.

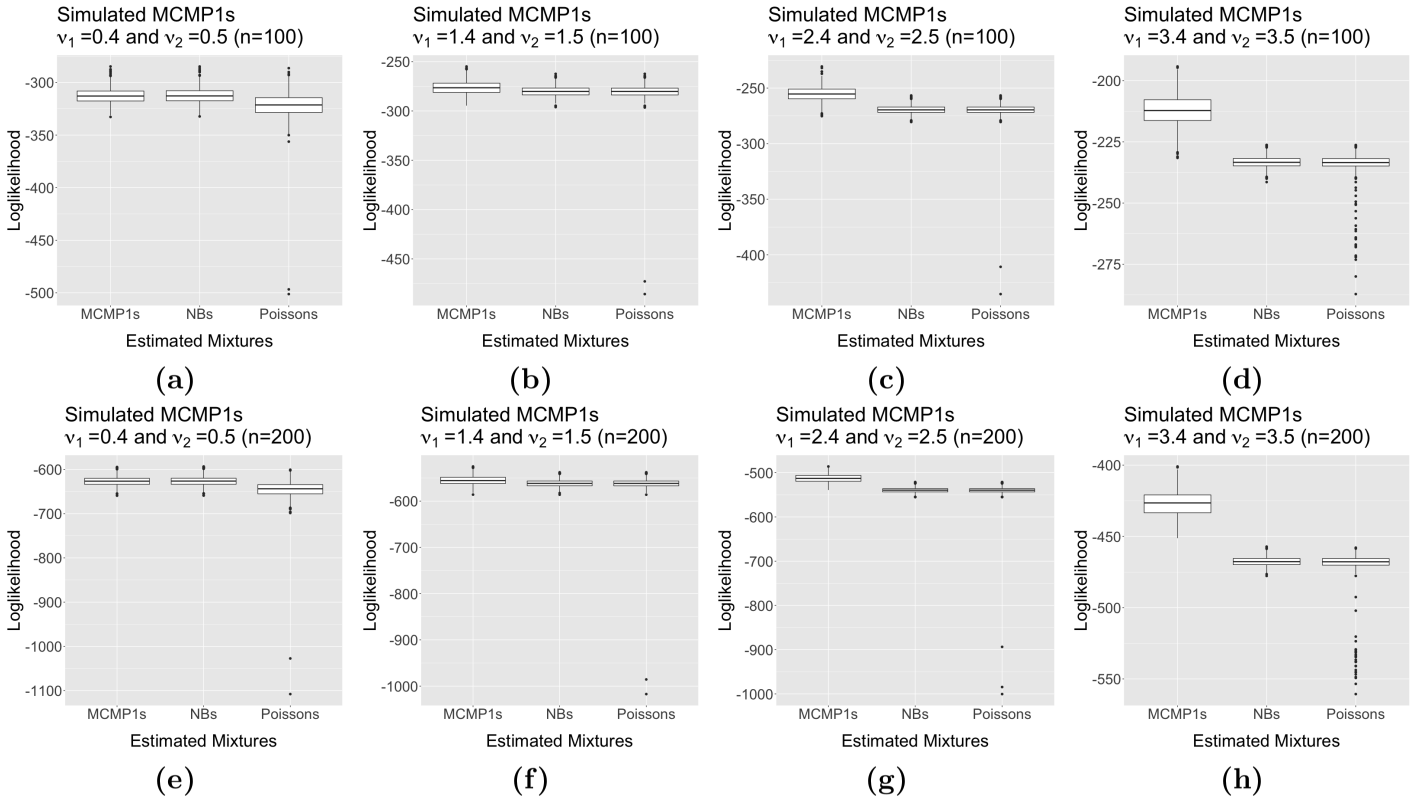


Figure S4: Comparative boxplots of loglikelihood values associated with two-component mixtures of MCMP1s, two-component mixtures of negative binomials (NBs), and two-component mixtures of Poissons fitted to the simulated data. Each boxplot summarizes the logliklelihood values for 1000 datasets generated from two-component mixtures of MCMP1s. Plots (a)–(d) summarize datasets of size $n = 100$ and plots (e)–(f) summarize datasets of size $n = 200$. The following dispersion parameters are used in the simulation: $(0.4, 0.5)$ for plots (a) and (e), $(1.4, 1.5)$ for plots (b) and (f), $(2.4, 2.5)$ for plots (c) and (g), and $(3.4, 3.5)$ for plots (d) and (h).

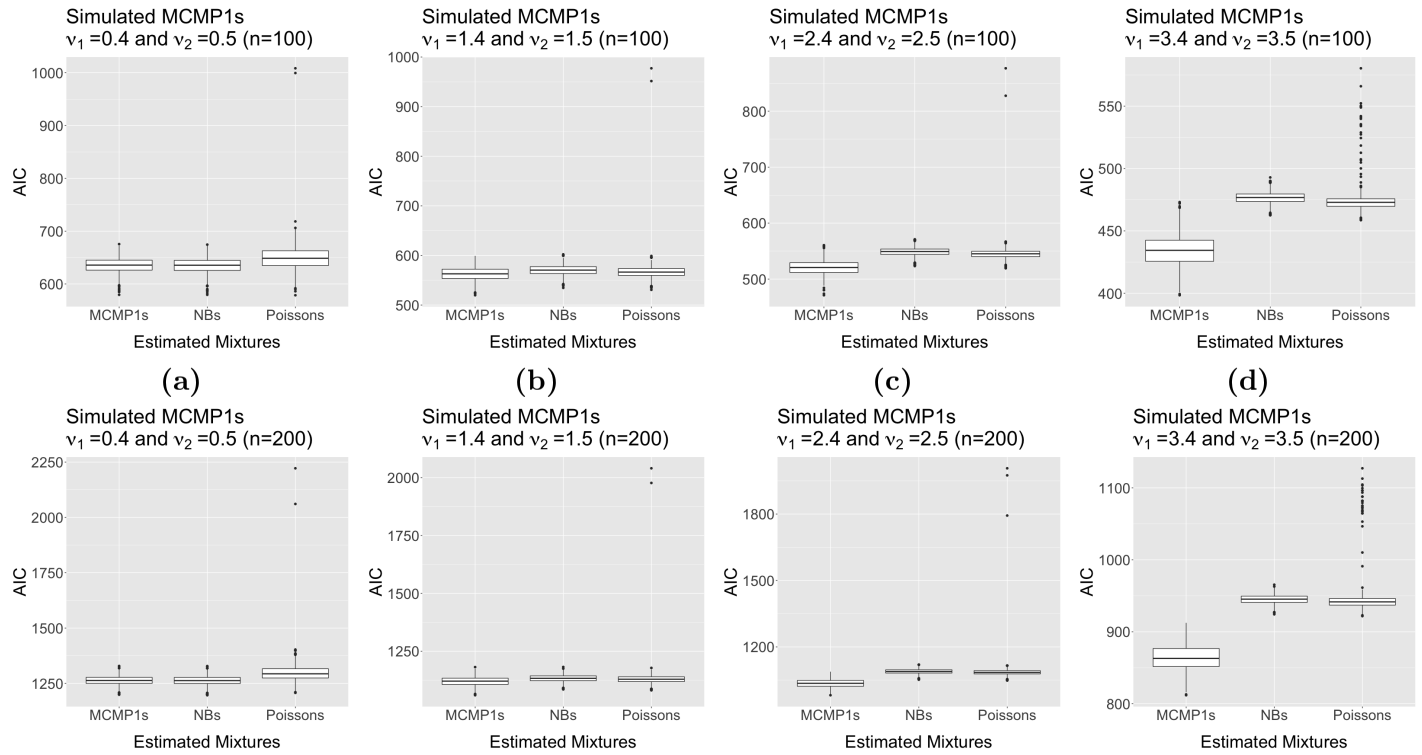


Figure S5: Comparative boxplots of AIC values associated with two-component mixtures of MCMP1s, two-component mixtures of negative binomials (NBs), and two-component mixtures of Poissons fitted to the simulated data. Each boxplot summarizes the AIC values for 1000 datasets generated from two-component mixtures of MCMP1s. Plots (a)–(d) summarize datasets of size $n = 100$ and plots (e)–(f) summarize datasets of size $n = 200$. The following dispersion parameters are used in the simulation: $(0.4, 0.5)$ for plots (a) and (e), $(1.4, 1.5)$ for plots (b) and (f), $(2.4, 2.5)$ for plots (c) and (g), and $(3.4, 3.5)$ for plots (d) and (h).

Table S1: Loglikelihood and AIC values for mixtures of Poissons, mixtures of negative binomials, and mixtures of MCMP1s fitted to the dog death data

	Loglikelihood			AIC		
m	Poissons	NBs	MCMP1s	Poissons	NBs	MCMP1s
1	-16992.8301	-16413.8141	-16193.0460	33987.6602	31396.7685	32390.0919
2	-15703.3672	-15696.3843	-15620.4848	31412.7344	31402.7686	31250.9696
3	-15703.3672	-15696.3811	-15562.9022	31416.7343	31408.7623	31141.8043
4	-15703.3679	-15694.2071	**-15555.3874**	31420.7343	31410.4141	**31132.7749**

2.10 Appendix C: R Code for EM Algorithm in Section 2.4

```
cmp.mixEM <- function(x, k=k,
                      mu=NULL, nu=NULL, Pi=NULL,
                      nu.star=seq(0.1,10,0.1),
                      eps=1e-6, maxit=1000){
  ## initial data
  x <- sort(x)      # count data
  n <- length(x)    # n sample size; k components

  ## initial parameters
  if (is.null(mu)) mu <- sort(as.vector(kmeans(x,k)$centers)) # k means
  if (is.null(nu)) nu <- rep(1,k)
  if (is.null(Pi)) Pi <- rep(1/k,k)


  ## initial observations in columns
  x.k <- matrix(nrow=n, ncol=k)
  for (i in 1:k) {
    x.k[,i] <- pmf(x,mu[i],nu[i])
  }

  ## observed loglikelihood
  obs.ll <- sum(log(rowSums(t(t(x.k)*Pi))))

  ## iteration to update the parameters
  iter <- 0
  dif <- 1

  # output
  out <- c(Pi,mu,nu,obs.ll,iter)
```

```
  while(iter < maxit && dif > eps){

    ## update parameters
    z.t <- t(t(x.k)*Pi) / rowSums(t(t(x.k)*Pi))
    pi <- colMeans(z.t)

    for (i in 1:k) {
      mu[i] <- weighted.mean(x,z.t[,i])
      nu[i] <- nu.fun(x,z.t[,i],mu[i],nu=nu.star)
    }

    ## update observations
    for (i in 1:k) {
      x.k[,i] <- pmf(x,mu[i],nu[i])
    }

    new.obs.ll <- sum(log(rowSums(t(t(x.k)*Pi))))
    dif <- abs(new.obs.ll-obs.ll)
    print(dif)

    obs.ll <- new.obs.ll
    print(iter)
    iter <- iter+1

    # output dataframe
    out <- rbind(out,c(Pi,mu,nu,new.obs.ll,iter))
  }

  colnames(out) <- c(paste("Pi",1:k,sep=""),
                     paste("mu",1:k,sep=""),
                     paste("nu",1:k,sep=""),"ll","iter")
  return(out)

}
```

Chapter 3 Finite Mixtures of Mean-Parameterized Conway-Maxwell-Poisson Regressions

3.1 Introduction

The Poisson distribution is, perhaps, the most widely used discrete distribution for the modeling of count data. However, the Poisson distribution has its mean equal to its variance (i.e., *equi-dispersion*), which limits its usage if the underlying data are, in fact, dispersed. The negative binomial distribution, with the form being parameterized as a Poisson-gamma mixture distribution such that it is characterized with the Poisson mean parameter and the gamma shape parameter (called the *type II negative binomial*, or *NB2*, by Hilbe 2011), is typically very effective in characterizing *over-dispersion* (i.e., variance is greater than its mean). Another popular and flexible discrete distribution is the Conway-Maxwell-Poisson distribution, which generalizes the Poisson distribution by introducing a dispersion parameter to address both data over-dispersion and *under-dispersion* (i.e., variance is less than its mean).

The CMP distribution was initially proposed by Conway and Maxwell (1961) as a queuing model with state-dependent service rates. It replaces the Poisson rate parameter with a generalized form: $\lambda = \mathrm{E}[Y^{\nu}] > 0$, where $\nu \geq 0$ is a dispersion parameter. The classic CMP distribution is a member of the exponential family, and a CMP random variable Y has probability mass function (pmf) given by

$$\mathrm{P}(Y = y \mid \lambda, \nu) = \frac{\lambda^{y}}{(y!)^{\nu}} \frac{1}{\mathcal{Z}(\lambda, \nu)}, \quad y = 0, 1, 2, \ldots, \tag{3.1}$$

where $\mathcal{Z}(\lambda, \nu) = \sum_{y=0}^{\infty} \frac{\lambda^{y}}{(y!)^{\nu}}$ is a normalizing constant that guarantees the pmf sums to unity. For the dispersion parameter, $\nu = 1$ yields a Poisson distribution, $\nu < 1$ indicates over-dispersion, and $\nu > 1$ indicates under-dispersion. The lack of proba-

bilistic and statistical characterization restricted mainstream application of the CMP for decades after its introduction. However, Shmueli et al. (2005) revived the CMP distribution with an extensive investigation into the distribution, which has made it an increasingly appealing model amongst statisticians and practitioners. Yet, unlike the Poisson distribution, the mean parameter λ neither has its ability to specify the distribution mean, nor locate the center of the distribution. Particularly, λ and ν do not have closed-form solutions when estimating the model parameters. A contemporary treatment about the CMP distribution is given in the text by Sellers (2023), which presents common acronyms for various CMP distributions that we use in the present paper.

Various reparameterizations have been employed to transform the CMP distribution into a form that is more convenient for application and to provide more direct interpretation. Specifically, these reparameterizations attempt to locate the center of the distribution. Guikema and Coffelt (2008) was the first work to reparameterize the CMP distribution by taking $\mu_\star = \lambda^{1/\nu}$ as the center, thus yielding an approximated CMP (ACMP). Ribeiro et al. (2020) used the approximated mean $\mu \approx \lambda^{1/\nu} - \frac{\nu-1}{2\nu}$ to reparameterize the CMP distribution (MCMP2), which is accurate when $\nu \leq 1$ and $\lambda^{1/\nu} > 10$ (Shmueli et al. 2005). Note that these parameterizations are based on approximations, and especially the latter is only applicable to over- or equi-dispersed data according to the accuracy of ν. Thus, this effectively prevents the ability of such a reparameterized CMP model to handle under-dispersion.

Another parameterization is the mean-parameterized CMP distribution (MCMP1) proposed by Huang (2017), which is characterized by the mean parameter $\mu \geq 0$ and dispersion parameter $\nu \geq 0$. The MCMP1 distribution has pmf

$$\mathrm{P}(Y = y \mid \mu, \nu) = \frac{\lambda(\mu,\nu)^y}{(y!)^\nu} \frac{1}{\mathcal{Z}(\lambda(\mu,\nu),\nu)}, \quad y = 0, 1, 2, \ldots, \tag{3.2}$$

where the rate parameter $\lambda(\mu, \nu)$ in the original CMP distribution is regarded as a

function of μ and ν, and is found by solving

$$\sum_{y=0}^{\infty}(y-\mu)\frac{\lambda(\mu,\nu)^y}{(y!)^\nu}=0. \tag{3.3}$$

The appeal of the MCMP1 is that it uses the true mean of the distribution for parameterization, so not only is the mean accurate for the entire parameter space *and* to denote the center, but the dispersion parameters are comparable across a variety of CMP distributions with the same μ (Huang 2017).

While observed count data often exhibit some degree of dispersion, the population from which the data are drawn may also consist of subpopulations. If there is, indeed, such a latent variable that can account for these subpopulations, then analyzing the data with a single, common distribution would not be appropriate. Instead, the use of a finite mixture model would provide a proper way to characterize the heterogeneity due to the latent subpopulations as well as identify distinct components for statistical analysis. We say that a random variable Y follows a mixture distribution with $m \in \mathbb{N}_+$ components if it has the mixture density

$$g(y;\boldsymbol{\Psi})=\sum_{j=1}^{m}\pi_j f_j(y;\boldsymbol{\theta}_j), \tag{3.4}$$

where the π_js are mixing proportions that satisfy $0 \leq \pi_j \leq 1$, $j = 1,\ldots,m$ and $\sum_{j=1}^{m}\pi_j = 1$. Here, the f_js are component-specific density (or mass) functions with $\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_j \subseteq \mathbb{R}^q$, where $\boldsymbol{\Theta}_j$ is open in $\mathbb{R}^q$. The mixture density in (3.4) is then parameterized by $\boldsymbol{\Psi} = \left(\pi_1,\ldots,\pi_{m-1},\boldsymbol{\theta}_1^{\mathrm{T}},\ldots,\boldsymbol{\theta}_m^{\mathrm{T}}\right)^{\mathrm{T}}$. Note that the f_j are typically from the same parametric distribution (e.g., Gaussian or Poisson), so the j index on f_j will often be suppressed in (3.4).

Finite mixture models have been developed for a vast array of data structures as well as applied to numerous diverse applications; see, for example, the texts by Lindsay (1995), McLachlan and Peel (2000), and Frühwirth-Schnatter (2006). One

class of mixture models that has a considerable body of literature is mixtures of linear regressions. The mixture of linear regressions problem has been extensively studied in the econometrics literature, where it was first introduced by Quandt (1972) as the *switching regressions*, or *switching regimes*, problem. Since then, many authors have addressed various inference considerations involving mixtures of linear regressions, as well as proposed flexible extensions to those models; cf. DeVeaux (1989), Viele and Tong (2002), Hurn et al. (2003), and Young and Hunter (2010). When the response variable is counts, then the components can be estimated via generalized linear models (GLMs). For example, mixtures of Poisson regressions have been applied in quality control to model the number of faults in a bolt of fabric (Aitken 1996) and in molecular biology to analyze high-throughput sequencing of RNA (Papastamoulis et al. 2016). Finite mixtures of binomial regressions were used to model the number of credits gained by freshmen during the first year at the School of Economics of the University of Florence (Grilli et al. 2015).

In mixture modeling of count data, it may be desirable to reflect varying degrees of dispersion across the components. Such a setting would suggest using a CMP distribution for the component distributions. However, there has been limited treatment of mixtures of CMP models. For example, Sur et al. (2015) developed a two-component mixture of truncated CMPs to analyze two datasets: the number of days spent in a hospital and Likert scale data for online ratings of a particular hotel. Zhan and Young (2023a) developed a general m-component mixture of MCMP1s, which was shown to be effective for analyzing a dog mortality dataset. However, neither of these models reflected dependency of the components on covariates. The present work fills that gap by developing mixtures of MCMP1 regressions.

The rest of this paper is organized as follows. In Sect. 3.2, we formally define the finite mixture of MCMCP1 regressions model. In Sect. 3.3, we provide the details for performing maximum likelihood estimation of this model via an expectation-

maximization (EM) algorithm (Dempster et al. 1977). In Sect. 3.4, we present the results from the simulation study for parameter estimation and model comparison. In Sect. 3.5, we apply our model to analyze data on the spread of a viral infection in potato plants by aphids. We end with a brief discussion in Sect. 3.6.

3.2 Mixtures of MCMP1 Regressions Model

In this section, we develop an m-component mixture of MCMP1 regressions model for the conditional distribution of $Y|\boldsymbol{X}$. Here, $Y \in \mathbb{N}$ is the discrete (count) response variable, and $\boldsymbol{X} = (X_1, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$ is a p-dimensional covariate vector. The mean parameters for the components are μ_j, $j = 1, \ldots, m$, which are modeled as a function of the covariates via a log link function (see Huang 2017 for the non-mixture setting) as

$$\mu_j = \exp(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_j), \tag{3.5}$$

where $\boldsymbol{x} = (x_0, x_1, \ldots, x_p)^{\mathrm{T}}$ and $\boldsymbol{\beta}_j = (\beta_{0j}, \beta_{1j}, \ldots, \beta_{pj})^{\mathrm{T}}$. Here, $x_0 = 1$ so as to allow for modeling with an intercept. Therefore, the β_{0j}s are the intercept of the jth component regression and the $\beta_{1j}, \ldots, \beta_{pj}$ are the coefficients for the corresponding covariates within the jth component regression.

Following the general mixture density given in (3.4), the m-component mixture of MCMP1 regressions model for $Y|\boldsymbol{X}$ has the mixture density

$$g(y; \boldsymbol{x}, \boldsymbol{\Psi}) = \sum_{j=1}^{m} \pi_j \frac{\lambda(\exp(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_j), \nu_j)^y}{(y!)^{\nu_j}} \frac{1}{\mathcal{Z}(\lambda(\exp(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_j), \nu_j), \nu_j)}, \tag{3.6}$$

where the parameter vector is

$$\boldsymbol{\Psi} = \left(\pi_1, \ldots, \pi_{m-1}, \boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_m^{\mathrm{T}}, \nu_1, \ldots, \nu_m\right)^{\mathrm{T}}.$$

Here, the π_js are, again, the mixing proportion for each component. The ν_js are

dispersion parameters, where it is assumed the data points in the same component follow this degree of dispersion.

3.3 EM Algorithm for Maximum Likelihood Estimation

In this section, we develop an EM algorithm to perform maximum likelihood estimation on the mixture of MCMP1 regressions model. Given a set of independent count observations $y_1, \ldots, y_n$ from the model in (3.6), measured with corresponding covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, the likelihood for the MCMP1 regression mixture model is

$$\mathcal{L}_o(\boldsymbol{\Psi}; \boldsymbol{y}, \mathbf{X}) = \prod_{i=1}^{n} \prod_{j=1}^{m} \pi_j \frac{\lambda(\exp(\boldsymbol{x_i}^{\mathrm{T}} \boldsymbol{\beta}_j), \nu_j)^{y_i}}{(y_i!)^{\nu_j}} \frac{1}{\mathcal{Z}(\lambda(\exp(\boldsymbol{x_i}^{\mathrm{T}} \boldsymbol{\beta}_j), \nu_j), \nu_j)}, \tag{3.7}$$

where the subscript "o" is used to denote the observed data. Here, $\boldsymbol{y}$ and $\mathbf{X}$ are used to denote, respectively, the y_is and $\boldsymbol{x}_i$s, $i = 1, \ldots, n$. As is typical with maximum likelihood estimation of mixture models, the likelihood in (3.7) is difficult to directly optimize. To make optimization tractable, we begin by noting that the observations $\boldsymbol{y}$ are considered incomplete, because their corresponding component labels are not observed, i.e., they are missing. To make the data complete, we define the indicator variables $Z_{ij} \sim \text{Bern}(\pi_j)$ to be the (unobserved) component label for observation i; specifically, $Z_{ij} = \mathtt{I}\{\text{if observation } i \text{ is from component } j\}$. Letting $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{im})^{\mathrm{T}}$, it follows that

$$\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n \overset{iid}{\sim} \text{Mult}_m(1, \{\pi_1, ..., \pi_m\}), \tag{3.8}$$

where $\text{Mult}_m(\cdot, \cdot)$ denotes the multinomial distribution with m categories. Note that there is a distinct use of Z as an indicator variable to represent the latent component membership, while $\mathcal{Z}(\cdot)$ is used as the normalizing constant in CMP pmfs. The use of "Z" is standard in both contexts [see McLachlan and Krishnan (2007) for the usage

in EM algorithms and Sellers (2023) for the usage in CMP modeling], which is why we have used slightly different formatting with these quantities to avoid an abuse of notation.

The complete data are, thus, $(y_i, \boldsymbol{x}_i, \boldsymbol{Z}_i)$, which yields the likelihood

$$\mathcal{L}_c(\boldsymbol{\Psi}) = \prod_{i=1}^{n}\prod_{j=1}^{m}\left\{\pi_j \frac{\lambda(\exp(\boldsymbol{x_i}^{\mathrm{T}}\boldsymbol{\beta}_j),\nu_j)^{y_i}}{(y_i!)^{\nu_j}} \frac{1}{\mathcal{Z}(\lambda(\exp(\boldsymbol{x_i}^{\mathrm{T}}\boldsymbol{\beta}_j),\nu_j),\nu_j)}\right\}^{Z_{ij}}. \tag{3.9}$$

Here, the subscript "c" is used to denote the complete data. The complete data loglikelihood is thus

$$\ell_c(\boldsymbol{\Psi}) = \sum_{i=1}^{n}\sum_{j=1}^{m} Z_{ij}\left\{\log\pi_j + \log\left(\frac{\lambda(\exp(\boldsymbol{x_i}^{\mathrm{T}}\boldsymbol{\beta}_j),\nu_j)^{y_i}}{(y_i!)^{\nu_j}\mathcal{Z}(\lambda(\exp(\boldsymbol{x_i}^{\mathrm{T}}\boldsymbol{\beta}_j),\nu_j),\nu_j)}\right)\right\}. \tag{3.10}$$

Since the $\boldsymbol{Z}_i$s are unknown, we use an EM algorithm to produce maximum likelihood estimation under the complete-data setup.

E-Step Given the parameters $\boldsymbol{\Psi}^{(t)}$ at the tth iteration, $t = 0, 1, 2, ...$, where $t = 0$ is used to denote the step where initial values are supplied, the expectation of $\ell_c(\boldsymbol{\Psi})$, conditioned on the observed data is computed as

$$\begin{aligned} Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(t)}) &= \mathbf{E}_{\boldsymbol{\Psi}^{(t)}}[\ell_c(\boldsymbol{\Psi})|\boldsymbol{y},\mathbf{X}] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij}^{(t)}\left\{\log\pi_j + \log\left(\frac{\lambda(\exp(\boldsymbol{x_i}^{\mathrm{T}}\boldsymbol{\beta}_j),\nu_j)^{y_i}}{(y_i!)^{\nu_j}\mathcal{Z}(\lambda(\exp(\boldsymbol{x_i}^{\mathrm{T}}\boldsymbol{\beta}_j),\nu_j),\nu_j)}\right)\right\}. \end{aligned} \tag{3.11}$$

The above expression depends on $z_{ij}^{(t)}$, which are referred to as *posterior membership probabilities*. These arise by noting that $\boldsymbol{Z}_{ij}$ is independent of $Y_{i'}$ for all $i \neq i'$. Since $\mathrm{E}_{\boldsymbol{\Psi}^{(t)}}$ is a linear functional, we may replace $\boldsymbol{Z}_{ij}$ by $\mathrm{E}_{\boldsymbol{\Psi}}[\boldsymbol{Z}_{ij} \mid Y_i = y_i]$, which when

provided the estimate $\boldsymbol{\Psi}^{(t)}$ yields

$$z_{ij}^{(t)} = \frac{\pi_j^{(t)} \left(\dfrac{\lambda(\exp(\boldsymbol{x_i}^{\mathrm{T}}\boldsymbol{\beta}_j^{(t)})_{ij}, \nu_j^{(t)})^{y_i}}{(y_i!)^{\nu_j^{(t)}} \mathcal{Z}(\lambda(\exp(\boldsymbol{x_i}^{\mathrm{T}}\boldsymbol{\beta}_j^{(t)})_{ij}, \nu_j^{(t)}), \nu_j^{(t)})} \right)}{\sum_{i=1}^{n}\sum_{k=1}^{m} \pi_k^{(t)} \left(\dfrac{\lambda(\exp(\boldsymbol{x_i}^{\mathrm{T}}\boldsymbol{\beta}_j^{(t)})_{ik}, \nu_k^{(t)})^{y_i}}{(y_i!)^{\nu_k^{(t)}} \mathcal{Z}(\lambda(\exp(\boldsymbol{x_i}^{\mathrm{T}}\boldsymbol{\beta}_j^{(t)})_{ik}, \nu_k^{(t)}), \nu_k^{(t)})} \right)}. \tag{3.12}$$

M-Step The maximization of $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(t)})$ with respect to $\boldsymbol{\Psi}$ gives the updated estimates $\boldsymbol{\Psi}^{(t+1)}$. First, through the direct use of a Lagrange multiplier, the updated mixing proportions are derived as

$$\frac{\partial Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(t)})}{\partial \pi_j} \stackrel{\text{set}}{=} 0 \Rightarrow \pi_j^{(t+1)} = \frac{\sum_{i=1}^{n} z_{ij}^{(t)}}{n}, \tag{3.13}$$

which is straightforward to compute as the average of the posterior membership probabilities of the observations belonging to the jth component. The updated regression parameters $\boldsymbol{\beta}_j$ for component j can be obtained by solving

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(t)})}{\partial \boldsymbol{\beta}_j} &= \frac{\partial Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(t)})}{\partial \lambda_{ij}} \frac{\partial \lambda_{ij}}{\partial \mu_{ij}} \frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}_j} \\ &= z_{ij} \frac{y_i - \mu_{ij}}{\lambda_{ij}} \boldsymbol{x}_i \exp(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_j) \stackrel{\text{set}}{=} \mathbf{0}, \end{aligned} \tag{3.14}$$

where, note, that $\mu_{ij} = \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j)$ and the chain rule has been invoked. Finally, the updated dispersion parameter ν_j for component j can be obtained by solving

$$\frac{\partial Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(t)})}{\partial \nu_j} = \sum_{i=1}^{n} z_{ij} \left\{ -\log(y_i!) + \mathrm{E}_{(\lambda_{ij}, \nu_j)}[\log(Y_i!)] \right\} \stackrel{\text{set}}{=} 0. \tag{3.15}$$

In the M-step, estimates for the $\boldsymbol{\beta}_j$s and ν_js require us to jointly solve (3.14) and (3.15). The `R` package `nloptr` (Johnson, 2014) is applied to solve the nonlinear

functions, where $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(t)})$ is the objective function for optimization. We set the $\boldsymbol{\beta}_j$s and ν_js, along with the original rate parameters λ_{ij}s, as unknowns. $\mu_{ij} = \exp(\boldsymbol{x_i}^{\mathrm{T}}\boldsymbol{\beta}_j)$ is set as the constraint functions. We must also specify bounds for the unknown parameters when using the `nloptr()` function, which we set as $\boldsymbol{\beta}_j \in \mathbb{R}^q$, $\nu_j \in [0.5, 10]$, and $\lambda_{ij} \in [0.1, 200]$. The bounded support for the λ_j and ν_j parameters are necessary for the `nloptr()` function to converge to reasonable estimates. Sur et al. (2015) show in their appendix that the CMP distribution tends to degenerate when ν is less than 0.5 or larger than 10. Also, $\lambda_{ij} \in [0.1, 200]$ are often reasonable for a CMP distribution as the count model is being fit to data that typically do not exhibit large count values. The gradients of the objective and constraint functions are formulated with respect to the $\boldsymbol{\beta}_j$s, ν_js, and λ_{ij}s separately. The global and local algorithm `NLOPT_LD_SLSQP` is chosen for the optimization process, which is used when maximizing the objective function $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(t)})$ to yield the updated $\boldsymbol{\beta}_j^{(t+1)}$s and $\nu_j^{(t+1)}$s.

The initial values impact the performance of EM algorithms when trying to find a maximum likelihood solution for a mixture model. Granted, this can vary greatly depending on the complexity of the model. Different strategies can be employed to find the best solution of the model; see, for example, Chapter 2 of McLachlan and Peel (2000). It is no different for the mixtures of MCMP1 regressions in this work. We adopted a straightforward strategy of using the $\boldsymbol{\beta}_j$s from mixtures of Poisson regressions as starting values for the $\boldsymbol{\beta}_j$s when estimating a mixture of MCMP1 regressions model. The initial $\boldsymbol{\beta}_j$s are the best among 100 fits of Poisson regression mixtures that were determined using the `R` package `flexmix` (Leisch 2004). This strategy makes sense because of the MCMP1 distribution being a generalization of the Poisson distribution. Generally, the estimated component means for a mixture of MCMP1 regressions will be close to the corresponding estimated component means if fitting a mixture of Poisson regressions. The initial mixing proportions are simply set at $1/m$ and the initial dispersion values are all set equal to 1.

The observed loglikelihood values at subsequent iterations are used to determine the convergence of our EM algorithm. The EM algorithm terminates when the criterion $\ell_o(\boldsymbol{\Psi}^{(t+1)}) - \ell_o(\boldsymbol{\Psi}^{(t)}) < \epsilon$ satisfied for some small fixed $\epsilon > 0$. Since the MCMP1 distribution does not change the CMP density, it is still from the exponential family (Huang 2017). Therefore the estimates always converge (Wu 1983).

3.4 Simulation Study

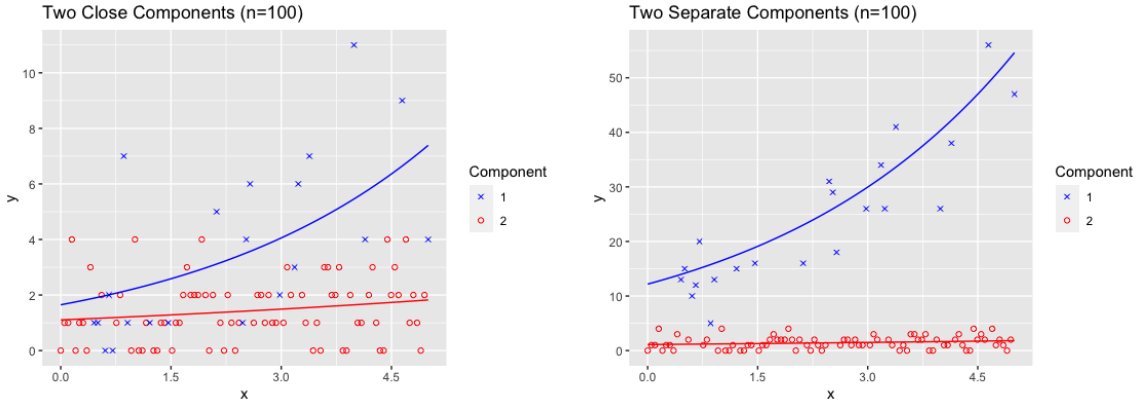
In this section, we use simulated data to evaluate the proposed mixture of MCMP1 regressions model and the EM algorithm. The discrete response variable is related to a continuous covariate and is simulated according to mixture of MCMP1 regressions. Models with two or three components are investigated, along with different sample sizes. The simulated data from the two-component model are also compared to estimates obtained from two-component mixtures of Poisson regressions and two-component mixtures of negative binomial regressions.

3.4.1 Parameter Estimates

For our simulation study, we only assume a single covariate x, which is set at equally-spaced values over the range $[0, 5]$. The response variable y is simulated from a two-component mixture of MCMP1 regressions using the parameters specified in Table 3.1 and a three-component mixture of MCMP1 regressions in Table 3.2. Parameter combinations were selected to yield two settings: one where data from the different components are close to each other and one where data from the different components are more separated. Datasets of sample sizes $n \in \{50, 100, 200\}$ for each mixture model were generated.

To help visualize a typical sample generated under the different mixtures of MCMP1 regressions models under consideration, we provide figures for a simulated dataset of size $n = 100$ along with the true component conditional means from these

models. Figure 3.1 shows a typical sample from the two-component models using the parameters given in Table 3.1. The sample points are overlaid with the true regression lines in the same color for two close components in Figure 3.1a and two well-separated separated components in Figure 3.1b. For the case of two close components, the regression parameters for the components are $\boldsymbol{\beta}_1 = (0.5, 0.3)^{\mathrm{T}}$ and $\boldsymbol{\beta}_2 = (0.1, 0.1)^{\mathrm{T}}$. The mixing proportion parameters are $(\pi, 1-\pi) = (0.3, 0.7)$, and the dispersion parameters are $(\nu_1, \nu_2) = (0.8, 1.2)$ for the two components. For the case of two separate components, the analogous parameters are $\boldsymbol{\beta}_1 = (2.5, 0.3)^{\mathrm{T}}$, $\boldsymbol{\beta}_2 = (0.1, 0.1)^{\mathrm{T}}$, $(\pi, 1-\pi) = (0.3, 0.7)$, and $(\nu_1, \nu_2) = (0.8, 1.2)$.

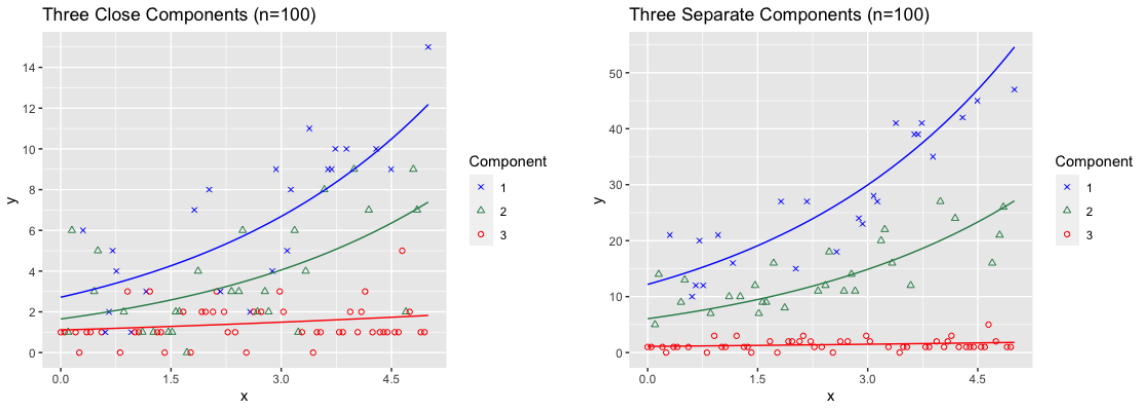


(a) **(b)**

Figure 3.1: Monte Carlo samples ($n = 100$) consisting of two components overlaid with the conditional mean lines using the true parameters

Figure 3.2 shows a typical sample from the three-component models using the parameters given in Table 3.2. The sample points from different components are overlaid with the true regression lines in the same color. Figure 3.2a shows the setting with three close components, where the regression parameters for the components are $\boldsymbol{\beta}_1 = (1, 0.3)^{\mathrm{T}}$, $\boldsymbol{\beta}_2 = (0.5, 0.3)^{\mathrm{T}}$, and $\boldsymbol{\beta}_3 = (0.1, 0.1)^{\mathrm{T}}$. The corresponding mixing proportion parameters are $(\pi_1, \pi_2, 1-\pi_1-\pi_2) = (0.2, 0.3, 0.5)$ and the dispersion parameters are $(\nu_1, \nu_2, \nu_3) = (0.8, 1, 1.2)$. Figure 3.2b shows the setting with three well-separated components. Here, the analogous parameters are

$\boldsymbol{\beta}_1 = (2.5, 0.3)^{\mathrm{T}}$, $\boldsymbol{\beta}_2 = (1.8, 0.3)^{\mathrm{T}}$, $\boldsymbol{\beta}_3 = (0.1, 0.1)^{\mathrm{T}}$, $(\pi_1, \pi_2, 1-\pi_1-\pi_2) = (0.2, 0.3, 0.5)$, and $(\nu_1, \nu_2, \nu_3) = (0.8, 1, 1.2)$.



(a) **(b)**

Figure 3.2: Monte Carlo samples ($n = 100$) consisting of three components overlaid with the conditional mean lines using the true parameters

When estimating the mixture models in this part of our study, we chose to use the true $\boldsymbol{\beta}_j$s as the initial values to start our EM algorithm. We always set the mixing proportions at $1/m$, and the dispersion parameters at 1. We then used the parameter estimates to assess the bias and root mean squared error (RMSE) for each of the estimators. These calculations were based on $M = 1000$ Monte Carlo samples. While sampling variability can make data generated from the underlying mixture models challenging to estimate when n is small, the overall performance of our approach is generally quite good in terms of the average biases and RMSEs, which we now look at more closely.

Tables 3.1 and 3.2 summarize the bias and RMSE values of the estimates obtained when fitting the corresponding mixture of MCMP1 regressions. In both tables, the biases for the mixing proportions are relatively small compared to their true parameters, which indicates that the proportion of contribution from each component to the overall mixture is being accurately captured. The bias values for all of the regression parameters in both settings are generally small, indicating overall good performance

when estimating the conditional mean structure. The dispersion estimates tend to have slightly larger biases and RMSEs, but still consistent on the order observed with the other parameters. The RMSE values decrease as n increases for all estimates in general. Overall, the results in this section indicate that our EM algorithm performs not bad at estimating the mixtures of MCMP1 regressions models used in this study.

Table 3.1: The average biases and RMSEs from $M = 1000$ datasets from two-component mixtures of MCMP1 regressions

$m = 2$ Close Components				$m = 2$ Separate Components			
Parameters	n	Bias	RMSE	Parameters	n	Bias	RMSE
$\beta_{01} = 0.5$	50	-0.0422	0.1024	$\beta_{01} = 2.5$	50	-0.0233	0.0412
	100	-0.0211	0.0641		100	0.0029	0.0274
	200	-0.0153	0.0460		200	0.0018	0.0187
$\beta_{11} = 0.3$	50	-0.0121	0.0313	$\beta_{11} = 0.3$	50	-0.0027	0.0048
	100	-0.0063	0.0206		100	0.0004	0.0032
	200	-0.0047	0.0151		200	0.0002	0.0022
$\beta_{02} = 0.1$	50	-0.1112	0.1666	$\beta_{02} = 0.1$	50	-0.0498	0.0936
	100	-0.1090	0.1410		100	-0.0395	0.0674
	200	-0.1019	0.1246		200	-0.0319	0.0508
$\beta_{12} = 0.1$	50	-0.0102	0.0183	$\beta_{12} = 0.1$	50	-0.0049	0.0105
	100	-0.0103	0.0136		100	-0.0036	0.0067
	200	-0.0098	0.0120		200	-0.0031	0.0051
$\pi_1 = 0.3$	50	0.0725	0.1356	$\pi_1 = 0.3$	50	-0.0022	0.0647
	100	0.0567	0.0970		100	-0.0039	0.0454
	200	0.0457	0.0779		200	-0.0021	0.0334
$\pi_2 = 0.7$	50	-0.0725	0.1356	$\pi_2 = 0.7$	50	0.0022	0.0647
	100	-0.0567	0.0970		100	0.0039	0.0454
	200	-0.0457	0.0779		200	0.0021	0.0334
$\nu_1 = 0.8$	50	0.2632	0.2990	$\nu_1 = 0.8$	50	0.2093	0.2097
	100	0.2297	0.2442		100	0.1989	0.1992
	200	0.2204	0.2279		200	0.1993	0.1994
$\nu_2 = 1.2$	50	0.0329	0.3167	$\nu_2 = 1.2$	50	-0.1118	0.1839
	100	0.0070	0.1892		100	-0.1336	0.1646
	200	-0.0108	0.1605		200	-0.1467	0.1994

Table 3.2: The average biases and RMSEs from $M = 1000$ datasets from three-component mixtures of MCMP1 regressions

m=3 Close Components				m=3 Separate Components			
Parameters	n	Bias	RMSE	Parameters	n	Bias	RMSE
$\beta_{01} = 1$	50	-0.0203	0.0949	$\beta_{01} = 2.5$	50	-0.0120	0.0441
	100	-0.0014	0.0665		100	0.0101	0.0348
	200	0.0061	0.0549		200	0.0109	0.0276
$\beta_{11} = 0.3$	50	-0.0042	0.0225	$\beta_{11} = 0.3$	50	-0.0013	0.0052
	100	0.0001	0.0160		100	0.0012	0.0041
	200	0.0018	0.0133		200	0.0013	0.0032
$\beta_{02} = 0.5$	50	-0.1069	0.1692	$\beta_{02} = 1.8$	50	0.0032	0.0492
	100	-0.0935	0.1456		100	0.0030	0.0349
	200	-0.0807	0.1209		200	0.0041	0.0259
$\beta_{12} = 0.3$	50	-0.0327	0.0505	$\beta_{12} = 0.3$	50	0.0006	0.0078
	100	-0.0294	0.0446		100	0.0005	0.0055
	200	-0.0254	0.0373		200	0.0007	0.0041
$\beta_{03} = 0.1$	50	-0.1110	0.1880	$\beta_{02} = 0.1$	50	-0.0558	0.1119
	100	-0.1007	0.1644		100	-0.0413	0.0795
	200	-0.1018	0.1452		200	-0.0366	0.0615
$\beta_{13} = 0.1$	50	-0.0096	0.0188	$\beta_{12} = 0.1$	50	-0.0055	0.0121
	100	-0.0092	0.0153		100	-0.0040	0.0079
	200	-0.0096	0.0135		200	-0.0036	0.0061
$\pi_1 = 0.2$	50	0.0534	0.1213	$\pi_1 = 0.2$	50	-0.0077	0.0643
	100	0.0350	0.0958		100	-0.0092	0.0448
	200	0.0293	0.0808		200	-0.0090	0.0321
$\pi_2 = 0.3$	50	0.0420	0.1599	$\pi_2 = 0.3$	50	0.0073	0.0737
	100	0.0445	0.1299		100	0.0052	0.0513
	200	0.0382	0.0999		200	0.0100	0.0382
$\pi_3 = 0.5$	50	-0.0954	0.1581	$\pi_3 = 0.5$	50	0.0005	0.0734
	100	-0.0796	0.1267		100	0.0040	0.0517
	200	-0.0675	0.0986		200	-0.0011	0.0361
$\nu_1 = 0.8$	50	0.2215	0.2373	$\nu_1 = 0.8$	50	0.2048	0.2055
	100	0.2029	0.2111		100	0.1961	0.1965
	200	0.1957	0.2012		200	0.1957	0.1960
$\nu_2 = 1$	50	0.1768	0.3028	$\nu_2 = 1$	50	-0.0014	0.0261
	100	0.1501	0.2551		100	-0.0015	0.0185
	200	0.1228	0.1956		200	-0.0021	0.0138
$\nu_3 = 1.2$	50	0.0583	0.3992	$\nu_3 = 1.2$	50	-0.0892	0.2300
	100	0.0196	0.3198		100	-0.1255	0.1795
	200	0.0057	0.2482		200	-0.1370	0.1623

3.4.2 Model Comparison

For the second part of our simulation study, we evaluate the performance of the mixture of MCMP1 regressions model when comparing with estimates obtained from mixtures of Poisson regressions and mixtures of negative binomial regressions. We simulated data from two-component mixtures of MCMP1 regressions with various dispersion parameters, and then fit the three different mixture models. The EM algorithm discussed in Sect. 3.3 is used for estimating the mixture of MCMP1 regressions model. The `flexmix()` function from the `flexmix` package is used to estimate the mixture of Poisson regressions model. We then wrote an EM algorithm using the `glm.nb()` function from the `MASS` package (Venables and Ripley 2002) to estimate the mixture of negative binomial regressions model. We estimated the mixture of MCMP1 regressions by initializing our EM algorithm at the true parameter values. For estimating the mixture of Poisson regressions model, we initially proceeded by finding the best fit from among 100 random starts based on the randomization routine underlying the `flexmix()` function, and then used those estimates as the initial values for the EM algorithms to estimate the other two mixture models. However, when also initializing the algorithm using the true regression parameters and mixing proportions used to generate the mixture of MCMP1 regressions data, this always converged to the best solution. We further initialized our mixture of negative binomial regressions EM algorithm the same way, but set the initial dispersion parameters equal to 1.

The data generated in this section have mixing proportion parameters $(\pi, 1-\pi) = (0.3, 0.7)$, and regression parameters $\boldsymbol{\beta}_1 = (2.5, 0.3)^{\mathrm{T}}$ and $\boldsymbol{\beta}_2 = (0.1, 0.1)^{\mathrm{T}}$. The two components are set as separate, and as shown in Figure 3.1b. In order to investigate how the dispersion parameters impact the model selection, we consider five cases with a variety of dispersion parameters assigned to the two components. The five cases have the dispersion parameters $(0.6, 0.9)$, $(0.8, 1.2)$, $(1.4, 1.5)$, $(2.4, 2.5)$, and

(3.4, 3.5). Case 1 consists of two over-dispersed components since both dispersion parameters are less than 1. Case 2 consists of one over-dispersed component and one under-dispersed component. Cases 3 to 5 consist of two under-dispersed components and have increasing degrees of under-dispersion as the parameters increase and are greater than 1. For each case, we generated $M = 1000$ datasets with the same sample sizes as in the previous study in Sect. 3.4.1.

The converged values of the observed loglikelihoods – denoted by $\ell_o^{(\infty)}$ – are obtained from each of the estimated mixture models. The proportion of times when the loglikelihood from the mixture of MCMP1 regressions fit is greater than that of the mixture of Poisson regressions fit, and is greater than that of the mixtures of negative binomial regressions fit, are reported separately in Table 3.3. Based on this metric, we see that the mixture of MCMP1 regressions is almost unanimously better than the mixture of Poisson regressions for all five cases. There are only a few datasets

Table 3.3: The proportion of the loglikelihood values from mixture of MCMP1 regressions fits (ℓ_{MCMP1s}) greater than that from mixture of Poisson regressions fits (ℓ_{Poissons}) or mixture of negative binomial regressions fits (ℓ_{NBs}) when the data were generated from two-component mixtures of MCMP1 regressions

Case (Dispersions)	n	$(\ell_{\text{MCMP1s}} > \ell_{\text{Poissons}})\%$	$(\ell_{\text{MCMP1s}} > \ell_{\text{NBs}})\%$
1 $(\nu_1 = 0.6, \nu_2 = 0.9)$	50	0.950	0.241
	100	0.965	0.072
	200	0.984	0.006
2 $(\nu_1 = 0.8, \nu_2 = 1.2)$	50	0.952	0.607
	100	0.982	0.438
	200	0.970	0.350
3 $(\nu_1 = 1.4, \nu_2 = 1.5)$	50	0.996	0.930
	100	1.000	0.967
	200	1.000	0.987
4 $(\nu_1 = 2.4, \nu_2 = 2.5)$	50	1.000	1.000
	100	1.000	1.000
	200	1.000	1.000
5 $(\nu_1 = 3.4, \nu_2 = 3.5)$	50	1.000	1.000
	100	1.000	1.000
	200	1.000	1.000

that yielded a mixture of Poisson regressions fit as better, which only occurred under Cases 1 and 2 when over-dispersion is present. When comparing to the mixture of negative binomial regressions fit, the mixture of MCMP1 regressions fit is better a strong majority of the time for all five cases. In Cases 1 and 2 when over-dispersion is present, there is a little more competitiveness between the two models, but as the amount of under-dispersion increases in each component, the mixture of MCMP1 regressions becomes unanimously the best fit.

Although the loglikelihood comparisons shows the performance of the mixture of MCMP1 regressions model, with notable superiority when the components are under-dispersed, the loglikelihoods from the three different mixture models are generally relatively close. This indicates that the three models are, perhaps, fairly comparable in most applications. As a further comparison, the Bayesian information criterion (BIC) values are calculated for each model. Recall that the BIC formula is $-2\ell_o^{(\infty)} + d\log(n)$, where d is the number of parameters in the model, and n is still the sample size. $d\log(n)$ forms the penalty term for model over-fitting. We can shift from a "best" model narrative, as we did with the loglikelihood comparisons, and apply the notion of BIC differences as introduced by Raftery (1995). The BIC difference provides a level of empirical support amongst candidate models. Formally, the BIC difference is defined as

$$\Delta\text{BIC}_i = \text{BIC}_i - \min_{i^* \in \mathcal{I}}(\text{BIC}_{i^*}),$$

where $\mathcal{I} = \{\text{Poisson}, \text{NB}, \text{MCMP1}\}$ is a set of the three count distributions considered for the components in the mixtures of regressions models being considered. Thus, ΔBIC_i is computed for each of the mixtures of Poisson regressions, mixtures of negative binomial (NB) regressions, and mixtures of MCMP1 regressions. We use $\Delta\text{BIC}_i \leq 10$, which is a threshold guided by Table 6 of Raftery (1995). In that table, the author states that $\Delta\text{BIC}_i > 10$ is indicative of "very strong" evidence in favor of the model with the minimum BIC. As ΔBIC_i decreases towards 0, the categories

indicate weaker evidence for distinguishing the "best" model between the two being compared.

Using the estimated models analyzed in Table 3.3, we then calculate the BIC for each model. The proportion of times $\Delta\text{BIC}_i \leq 10$ for each candidate model is shown in Table 3.4. A larger proportion of $\Delta\text{BIC}_i \leq 10$ demonstrates that candidate model i is comparable to the one that is considered "best" based on the BIC values. Since the components in a mixture of Poisson regressions model do not have a dispersion parameter to be estimated, this gives the mixture of Poisson regressions an advantage with BIC because the penalty will be smaller (due to fewer parameters) relative to the penalty for the mixture of MCMP1 regressions and mixture of negative binomial regressions. Thus, it is not surprising that a noticeably smaller penalty term is easy to achieve for the mixture of Poisson regressions BIC value given that the loglikelihood values from the three candidate models are all very close. When one or both of

Table 3.4: The proportion of times when $\Delta\text{BIC}_i < 10$ for each of the candidate mixture of regressions models when $M = 1000$ datasets are generated from a two-component mixture of MCMP1 regressions model

Case (Dispersions)	n	$\#(\Delta\text{BIC}_i \leq 10)/M$		
		Poisson	NB	MCMP1
1 $(\nu_1 = 0.6, \nu_2 = 0.9)$	50	0.999	0.999	0.948
	100	0.989	1.000	0.840
	200	0.961	0.951	0.049
2 $(\nu_1 = 0.8, \nu_2 = 1.2)$	50	1.000	0.999	0.955
	100	1.000	1.000	0.981
	200	0.999	0.539	0.161
3 $(\nu_1 = 1.4, \nu_2 = 1.5)$	50	1.000	1.000	0.996
	100	1.000	1.000	1.000
	200	1.000	0.003	0.627
4 $(\nu_1 = 2.4, \nu_2 = 2.5)$	50	1.000	1.000	1.000
	100	1.000	1.000	1.000
	200	1.000	0.000	1.000
5 $(\nu_1 = 3.4, \nu_2 = 3.5)$	50	1.000	0.999	1.000
	100	1.000	1.000	1.000
	200	1.000	0.000	1.000

the two components are over-dispersed (i.e., Cases 1 and 2), the mixtures of negative binomial regressions are almost consistently the best or comparable to the best fitting model. Based on $\Delta\text{BIC}_i \leq 10$, we see that the mixtures of MCMP1 regressions are fairly competitive. The advantage of the mixture of MCMP1 regressions becomes more pronounced as the two components are more under-dispersed. In particular, Cases 3 through 5 have increasing amounts of under-dispersion, resulting in the mixture of MCMP1 regressions being almost unanimously the best or comparable to the best fitting model, whereas the other two mixture models show relatively little competitiveness, especially when $n = 200$. Overall, these results are consistent with the analysis provided about the loglikelihoods from these estimated models.

3.5 Application: Aphids Data

Turner (2000) presented and analyzed data from an experiment designed to investigate how green peach aphids, a highly-infectious insect for plants, spread a viral infection among potato plants. The number of infected plants and the number of aphids released from the flight chamber were recorded. The response is the number of infected plants, which ranges from 0 to 27. The primary covariate is the number of aphids released, which ranges from 0 to 320. The experiment involved a total of $n = 51$ batches of aphids that were released. Turner (2000) thoroughly analyzed these data using a two-component mixture of linear regressions, and discussed other relevant inferential procedures in that work. Other works have also analyzed these data in the context of mixtures of linear regressions research; cf. Grün and Leisch (2008), Young and Hunter (2010), and Kasahara and Shimotsu (2015). However, the response variable is actually a count, so these data should more appropriately be analyzed using a mixture of count regressions model. Thus, we will turn to investigating the fits from mixtures of MCMP1 regressions as well as mixtures of Poisson regressions and mixtures of negative binomial regressions.

For each of these mixtures of count regression models, we consider $m \in \{1, 2, 3\}$ for the number of components. As in Sect. 3.4.2, we use the `flexmix()` function to estimate the mixtures of Poisson regressions. The fit with the smallest BIC value is chosen to represent the model among 100 random starts. The EM algorithm we wrote using the `glm.nb()` function was again used to estimate the mixtures of negative binomial regressions, while the EM algorithm introduced in Sects. 3.3 and 3.4 was used to estimate the mixtures of MCMP1 regressions. We again use the estimates from the mixture of Poisson regressions as the starting values for the regression parameters in our EM algorithms. The initial mixing proportions are all set equal to $1/m$ and the initial dispersions are all set equal to 1.

The BIC values from all of these fits are shown in Table 3.5. In each model category, two components are identified as the best given they have the smallest BIC value in the respective column. The two-component mixture of Poisson regressions has a smaller BIC value than the other two models with two components. Those other two models have very similar BIC values with only a difference of 0.3997 between them. The results are consistent with the simulation results in Sect. 3.4.2 in that while the mixture of MCMP1 regressions does not have the smallest BIC value, it does have a relatively small BIC difference of 7.7835, indicating that it is still a comparable model relative to the two-component mixture of Poisson regressions. But in spite of that, the mixture of MCMP1 regressions model affords us the flexibility

Table 3.5: BIC values for mixtures of Poisson regressions, mixtures of negative binomial regressions, and mixtures of MCMP1 regressions when those models are fit to the aphids data

	BIC		
m	Poissons	NBs	MCMP1s
1	403.3398	283.9228	401.2514
2	275.0845	282.4683	282.8680
3	285.0741	298.1957	368.1247

of having components that characterize deviations, no matter how small, from the equi-dispersion assumption that underlies the components of a mixture of Poisson regressions.

The estimates and the estimated standard errors for the parameters from each of the two-component models are reported in Table 3.6. The standard errors are estimated using a parametric bootstrap with $B = 1000$ bootstrap samples generated from each mixture model fit. The three mixture models yield similar mixing proportion estimates: $(0.4638, 0.5362)$ for the mixture of Poisson regressions, $(0.4550, 0.5450)$ for the mixture of negative binomial regressions, and $(0.4623, 0.5377)$ for the mixture of MCMP1 rgressions. The estimated standard errors for the mixing proportion estimates are also very similar. Moreover, the regression parameter estimates are similar from the three mixtures of regressions fits and their estimated standard errors are similar in magnitude across the three estimated models. Noticeably, one component from the mixture of negative binomial regressions fit has an extremely large dispersion estimate. Since our EM algorithm makes use of the `glm.nb()` function, this uses the Poisson-gamma mixture representation of the negative binomial. Specifically, as

Table 3.6: The estimates and corresponding estimated standard errors for the aphids data fit using the two-component mixture of Poisson regressions, mixture of negative binomial (NB) regressions, and mixture of MCMP1 regressions

	Poissons		NBs		MCMP1s	
Par.	Estimate	$\widehat{\text{SE}}$	Estimate	$\widehat{\text{SE}}$	Estimate	$\widehat{\text{SE}}$
π_1	0.4638	0.0367	0.4550	0.0794	0.4623	0.0752
π_2	0.5362	0.0367	0.5450	0.0794	0.5377	0.0752
β_{01}	1.7167	0.1365	1.7289	0.1886	1.7159	0.0482
β_{11}	0.0043	0.0006	0.0043	0.0018	0.0043	0.0001
β_{02}	-0.1081	0.3692	-0.0403	1.2629	-0.0859	0.0973
β_{12}	0.0019	0.0021	0.0018	0.0036	0.0020	0.0002
ν_1	-	-	146444	89974.2	1.0004	0.0270
ν_2	-	-	4.2248	9007.9	0.9564	0.2646
$\ell_o^{(\infty)}$	-127.7127		-127.4727		-127.6725	

discussed on page 3 of Hilbe (2011), there is an indirect relationship between the gamma shape parameter and the degree of over-dispersion in the data. So this means that the negative binomial effectively becomes a Poisson distribution as the dispersion parameter goes to infinity. Moreover, the corresponding estimated standard errors for the estimated dispersion for each component are extremely large. As a comparison, the dispersion estimates from mixture of MCMP1 regressions and the corresponding estimated standard errors have values that are of a much more reasonable magnitude.

The loglikelihoods for each model's fit are also provided in Table 3.6. All three models have very similar loglikelihood values, with the mixture of Poisson regressions fit technically being the "worst" because it is the smallest. However, as shown in Table 3.5, the two-component mixture of Poisson regressions has the smallest BIC value because of the smaller penalty used in the calculation. Hence, it would be the "best" as noted earlier. Regardless, this analysis shows that the mixture of MCMP1 regressions model is a reasonable competitor. Even though the dispersion parameter estimates most likely suggest that equi-dispersion is a tenable assumption in each component, the model still gives the flexibility to capture reasonable degrees of under-/over-dispersion for the individual components.

Figure 3.3 is a scatterplot of the aphids data with the estimated two-component mixture of MCMP1 regressions model overlaid. Different colors and plotting symbols are used to denote assignment of a point to a given component based on their maximum posterior membership probability. The conditional means plotted in Figure 3.3 are very similar to those estimated for the other two models in Table 3.6, so those are not overlaid on this figure. We see that the two MCMP1 regression models appear to be reasonable models for the respective components, and that they are adequately capturing the seemingly two different trends in this experiment. Note that this supports the possible explanation provided by Turner (2000), who stated that "some of the batches of aphids consisted of insects that had passed their 'maiden' phase. After

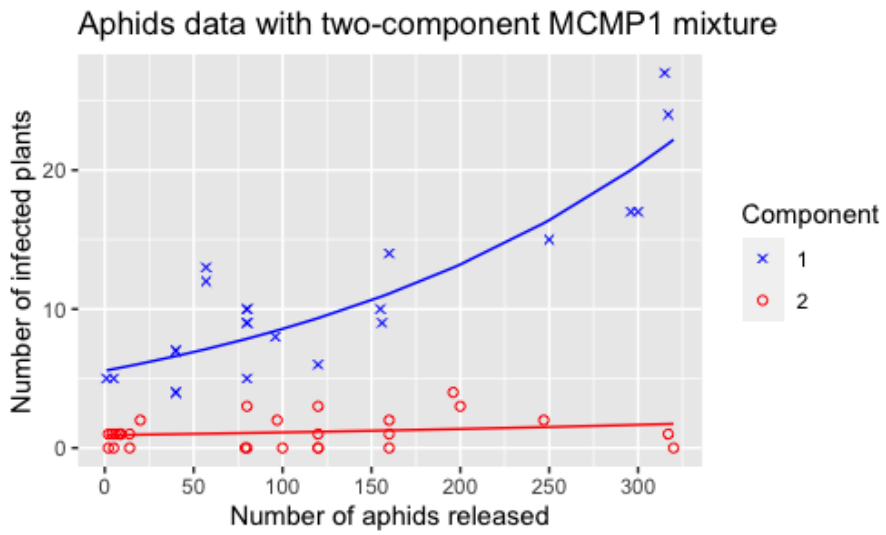


Figure 3.3: Scatterplot of the aphids data overlaid with the conditional mean lines estimated for the two-component mixture of MCMP1 regressions model

the maiden phase, aphids tend to settle on the first acceptable food host that they encounter, leading to low or zero levels of transmission of virus." Thus, the second component with practically zero slope would likely be those aphids that had passed their 'maiden' phase.

3.6 Discussion

In this paper, we motivated and developed the estimation of finite mixtures of MCMP1 regressions models. Our work contributes to the recent modeling developments involving the MCMP1, which was introduced by Huang (2017). Our work also contributes to the limited literature on mixtures of CMP models, which is primarily the works by Sur et al. (2015) and Zhan and Young (2023a). Both papers only treat univariate data and do not include modeling with covariates, which allows for much greater flexibility in modeling and understanding count datasets.

Maximum likelihood estimation for the mixture of MCMP1 regressions model was performed using an EM algorithm that we developed. The excellent performance of this algorithm was demonstrated via the simulation study in Sect. 3.4.1. An

additional model comparison study was performed in Sect. 3.4.2, where mixtures of Poisson regressions and mixtures of negative binomial regressions were considered as other candidate models. Overall, these results showed that the mixture of MCMP1 regressions tends to be a competitive model and, especially, the more practical model when data from a given component is under-dispersed. The performance of the mixture of MCMP1 regressions model was also demonstrated on the aphids data of Turner (2000).

One extension to our model that can provide additional flexibility is to further apply a GLM framework when modeling each of the dispersion and mixing proportion parameters as functions of covariates. This could be done by using a log link for ν_j and a logit link for π_j. One issue here will be to investigate identifiability of such a generalization to our mixture of MCMP1 regressions model.

3.7 Appendix A: R Code for EM Algorithm in Section 3.3

```
cmp.mixEMReg <- function(y, x, k=k,
                         beta=NULL, nu=NULL, Pi=NULL,
                         eps=1e-3, maxit=1000){

  X <- cbind(1,x)        # covariates

  n <- length(y)  # n: number of observations # k: number of components
  q <- ncol(X)    # q: column number of beta's

  ## initial beta's from Poisson mixtures
  if (is.null(beta)) {
    out.pois <- pois.mixReg.bestfit(y,x,k,maxr=100)
    beta <- matrix(out.pois[c("beta01","beta11","beta02","beta12")],
                   nrow=q,ncol=k)
  }

  ## initial Pi's and nu's
  if (is.null(nu)) nu <- rep(1,k)
  if (is.null(Pi)) Pi <- rep(1/k,k)

  ## initial observations
```

```
x.beta <- X %*% beta
mu0 <- exp(X %*% beta)
lambda0 <- matrix(nrow=n, ncol=k)
y.k <- matrix(nrow=n, ncol=k)
for (i in 1:n) {
  for (j in 1:k) {
    lambda0[i,j] <- lambda.fun(mu0[i,j],nu=nu[j],ylim=150)
    y.k[i,j] <- pmf(y[i], mu=mu0[i,j], nu=nu[j])
  }
}

## initial observed loglikelihood
obs.ll <- sum(log(rowSums(t(t(y.k)*Pi))))

## define posterior probabilities
z.t <- t(t(y.k)*Pi) / rowSums(t(t(y.k)*Pi))

## initial estimates summary
param <- c(c(beta),c(lambda0),nu)

#########################################################
### Define the functions for using nloptr package  ###
#########################################################

#####################################################
## objective function to maximize Q (i.e. minimize -Q)
Q.f <- function(param, k=k, y,X,z.t){

  q <- ncol(X)
  n <- length(y)

  ## parameters
  beta <- param[1:(q*k)]
  beta <- matrix(beta, nrow=q, ncol=k)

  lambda <- param[(q*k+1):(q*k+n*k)]
  lambda <- matrix(lambda, nrow=n, ncol=k)

  nu <- param[(q*k+n*k+1):(q*k+n*k+k)]

  ## values in Q
  nu.lfactorial <- matrix(nrow=n, ncol=k)
  ZZ <- matrix(nrow=n, ncol=k)
  for (i in 1:n) {
    for (j in 1:k) {
```

```
      nu.lfactorial[i,j] <- nu[j] * lgamma(y[i]+1)
      ZZ[i,j] <- Z(lambda[i,j],nu[j])
    }
  }

  Q <- sum(t(z.t) * log(Pi)) + sum(z.t * ( y * log(lambda)  -
       nu.lfactorial - log(ZZ)))

  return(-Q)

}

#################################################
## gradients of the objective function
Q.g <- function(param, k=k, y,X,z.t){

  q <- ncol(X)
  n <- length(y)

  ## parameters
  beta <- param[1:(q*k)]
  beta <- matrix(beta, nrow=q, ncol=k)

  lambda <- param[(q*k+1):(q*k+n*k)]
  lambda <- matrix(lambda, nrow=n, ncol=k)

  nu <- param[(q*k+n*k+1):(q*k+n*k+k)]
  mu <- exp(X %*% beta)

  ## values in gradients of Q
  mean_logfacy <- matrix(nrow=n, ncol=k)
  for (i in 1:n) {
    for (j in 1:k) {
      mean_logfacy[i,j] <- mean_logfactorialy.fun(lambda[i,j],nu[j])
    }
  }

  beta.grad <- rep(0,q*k)
  lambda.grad <- (z.t * y - z.t * mu)/lambda
  nu.grad <- colSums(z.t * (-lgamma(y+1)) + z.t *  mean_logfacy)

  return(-c(beta.grad,c(lambda.grad),nu.grad))

}
```

```
#####################################################
## equality constraint function
mu.con <- function(param, k=k, y,X,z.t){

  q <- ncol(X)
  n <- length(y)

  ## parameters
  beta <- param[1:(q*k)]
  beta <- matrix(beta, nrow=q, ncol=k)

  lambda <- param[(q*k+1):(q*k+n*k)]
  lambda <- matrix(lambda, nrow=n, ncol=k)

  nu <- param[(q*k+n*k+1):(q*k+n*k+k)]

  mu <- matrix(nrow=n, ncol=k)
  for (i in 1:n) {
    for (j in 1:k) {
      mu[i,j] <- mean.fun(lambda[i,j],nu[j],maxy=100,eps=1e-6)
    }
  }

  x.beta <- X %*% beta

  return(c(exp(x.beta)-mu))

}

#####################################################
## gradients of equality constraint function
mu.con.g <- function(param, k=k, y,X,z.t){

  q <- ncol(X)
  n <- length(y)

  ## parameters
  beta <- param[1:(q*k)]
  beta <- matrix(beta, nrow=q, ncol=k)

  lambda <- param[(q*k+1):(q*k+n*k)]
  lambda <- matrix(lambda, nrow=n, ncol=k)

  nu <- param[(q*k+n*k+1):(q*k+n*k+k)]
```

```
  ## beta gradients of constraint
  mu <- exp(X %*% beta)

  grad.beta <- t(X) %*% diag(mu[,1])
  if (k > 1) {
    for (i in 2:k) {
      grad.beta <- bdiag(grad.beta, t(X) %*% diag(mu[,i]) )
    }
  }

  ## lambda gradients of constraint
  V <- matrix(nrow=n, ncol=k)
  for (i in 1:n) {
    for (j in 1:k) {
      V[i,j] <- var.fun(lambda[i,j],nu[j],maxy=150,eps=1e-6)
    }
  }

  grad.lambda <- diag(V[,1]/lambda[,1])
  if (k > 1) {
    for (i in 2:k) {
      grad.lambda <- bdiag(grad.lambda, diag(V[,i]/lambda[,i]) )
    }
  }

  ## nu gradients of constraint
  grad.nu.m <- matrix(nrow=n, ncol=k)
  for (i in 1:n) {
    for (j in 1:k) {
      grad.nu.m[i,j] <-
  mean_ylogfactorialy.fun(lambda[i,j],nu[j],maxy=150,eps=1e-6)  -
  mu[i,j]*mean_logfactorialy.fun(lambda[i,j],nu[j],maxy=150,eps=1e-6)
    }
  }

  grad.nu <- grad.nu.m[,1]
  if (k > 1){
    for (i in 2:k) {
      grad.nu <- bdiag(grad.nu, grad.nu.m[,i] )
    }
  }

  return(as.matrix(cbind(t(grad.beta), grad.lambda, grad.nu)))

}
```

```
#####################################################################
## nloptr package to solve for the estimates by minimizing -Q ##
#####################################################################

#################################################
# define the upper and lower bounds for the parameters

## lower bounds
beta_l <- rep(-Inf, q*k)
lambda_l <- rep(0.1, n*k)
nu_l <- rep(0.5, k)

param_l <- c(beta_l, lambda_l, nu_l)

## upper bounds
beta_u <- rep(Inf, q*k)
lambda_u <- rep(200, n*k)
nu_u <- rep(10, k)

param_u <- c(beta_u, lambda_u, nu_u)
#################################################

#########################################################
################### EM algorithm ####################
#########################################################

## iteration starts
iter <- 0
dif <- 1

## output summary
out <- c(iter, obs.ll, Pi, c(beta), nu)

## iteration for EM algorithm
while(iter < maxit && dif > eps){

  ## nloptr to minimize -Q
  fit <- nloptr::nloptr(x0 = param,
                        eval_f = Q.f,
                        eval_grad_f = Q.g,
                        lb = param_l,
                        ub = param_u,
                        eval_g_eq = mu.con,
                        eval_jac_g_eq = mu.con.g,
```

```
                          opts = list("algorithm"= "NLOPT_LD_SLSQP",
                                      "maxeval" = 1000,
                                    "local_opts" = list("algorithm" = "NLOPT_LD_SLSQP",
                                    "xtol_rel" = 0.001)),k=k,y=y,X=X,z.t=z.t)

  ## update estimates
  param <- fit$solution
  beta <- param[1:(q*k)]
  beta <- matrix(beta, nrow=q, ncol=k)
  nu <- param[(q*k+n*k+1):(q*k+n*k+k)]

  ## update observations
  x.beta <- X %*% beta
  mu0 <- exp(X %*% beta)
  y.k <- matrix(nrow=n, ncol=k)
  for (i in 1:n) {
    for (j in 1:k) {
      y.k[i,j] <- pmf(y[i], mu=mu0[i,j], nu=nu[j])
    }
  }

  ## update loglikelihood
  new.obs.ll <- sum(log(rowSums(t(t(y.k)*Pi))))

  ## update posterior probabilites
  z.t <- t(t(y.k)*Pi) / rowSums(t(t(y.k)*Pi))

  ## update mixing proportions
  Pi <- colMeans(z.t)

  ## print
  iter <- iter+1
  print(iter)
  dif <- abs(new.obs.ll-obs.ll)
  print(dif)

  obs.ll <- new.obs.ll    # iteration ends

  ## output dataframe
  out <- rbind(out,c(iter, obs.ll, Pi, c(beta), nu))
}

colnames(out) <- c("iter", "ll",
                   paste("Pi",1:k,sep=""),
                   paste(rep(paste("beta",0:(q-1),sep=""),k),
```

```
                        rep(1:k,each=q),sep=""),
                  paste("nu",1:k,sep=""))

  return(out)

}
```

Chapter 4 Conclusions, Discussions, and Future Research

4.1 Conclusions

The Conway-Maxwell-Poisson distribution, as a discrete distribution, has been extensively studied for its unique ability to characterize both over-dispersion and under-dispersion in data. Equivalent to the original distribution, the mean-parameterized Conway-Maxwell-Poisson distribution (Huang 2017) offers convenience in interpreting data by noting the distribution's center through the mean parameter. Additionally, it provides insight into the degree of variation in observations around the mean using the dispersion parameter. In this dissertation, within the context of heterogeneous data requiring a mixture model, models along with estimation methods for mixtures of mean-parameterized Conway-Maxwell-Poisson distributions or regressions were successfully developed.

In Chapter 2, a finite mixture model comprising mean-parameterized Conway-Maxwell-Poisson distributions was proposed for univariate data. The EM algorithm was constructed for maximum likelihood estimation of the model. For the simulation study, replicated samples were generated from mixtures of two or three components of Conway-Maxwell-Poisson distributions. The parameter estimates from the simulation study demonstrated the model's validity by assessing biases and root mean squared errors (RMSEs). The model selection results in the simulation study showed that mixtures of mean-parameterized Conway-Maxwell-Poisson distributions and mixtures of negative binomials outperformed mixtures of Poissons for data with dispersions. Notably, mixtures of mean-parameterized Conway-Maxwell-Poisson distributions had the advantage over both Poisson mixtures and negative binomial mixtures in addressing increased levels of under-dispersion. The analysis of dog mortality data highlighted the capability of mean-parameterized Conway-Maxwell-Poisson mixtures

to identify a third component among the ages at death, and that was beyond the capabilities of Poisson mixtures and negative binomial mixtures. The estimates along with parametric and non-parametric bootstrap standard errors were provided for the dog mortality data.

In Chapter 3, within the framework of generalized linear models, a model for mixtures of mean-parameterized Conway-Maxwell-Poisson regressions was constructed. In this model, each response'mean was assumed to have a log-linear relationship with its covariates. Maximum likelihood estimation was conducted via the EM algorithm, with the `R` package `nloptr` employed to find solutions in the maximization step. The mixture responses with corresponding covariates were simulated from the mixtures of two or three mean-parameterized Conway-Maxwell-Poisson components. Parameter estimates from the simulation study indicated that the regression model performed quite well. The model comparison in the simulation study suggested that the three regression models—mixture of mean-parameterized Conway-Maxwell-Poisson regressions, mixture of Poisson regressions, and mixture of negative binomial regressions—were feasibly comparable for application to dispersed data. The aphids data was fitted using the three different mixtures of regressions models, and the estimates, along with its parametric bootstrap standard errors, were provided. Notably, the mixture of mean-parameterized Conway-Maxwell-Poisson regressions yielded much more reasonable dispersion estimates and standard errors, compared with the mixture of negative binomial regressions.

In summary, the mixture of mean-parameterized Conway-Maxwell-Poisson (regressions) models presented in this dissertation are capable to characterize component dispersions in data. These models offer valuable applications in specific scenarios, enriching the computational toolbox of finite mixture models for discrete data, alongside Poisson (regression) mixtures, negative binomial (regression) mixtures, and other models.

4.2 Discussions

4.2.1 Normalizing Constant and Mean

The Conway-Maxwell-Poisson distribution employs a dispersion parameter to adjust the Poisson distribution, forming more variability for over-dispersion or less variability for under-dispersion. To ensure that the probability theory is mathematically satisfied, a normalizing constant, $\mathcal{Z}(\lambda, \nu)$, is introduced in the Conway-Maxwell-Poisson distribution (1.1) to guarantee that the probabilities of all possible discrete values add up to unity.

A concern revolves around how the values of the normalizing constant $\mathcal{Z}(\lambda, \nu)$ behave throughout the entire parameter space (λ, ν). Figure 4.1 presents a surface plot of $\mathcal{Z}(\lambda, \nu)$ concerning $\lambda \in [1, 50]$ and $\nu \in [0, 5]$, with $\mathcal{Z}$ values being cut off at 10^5

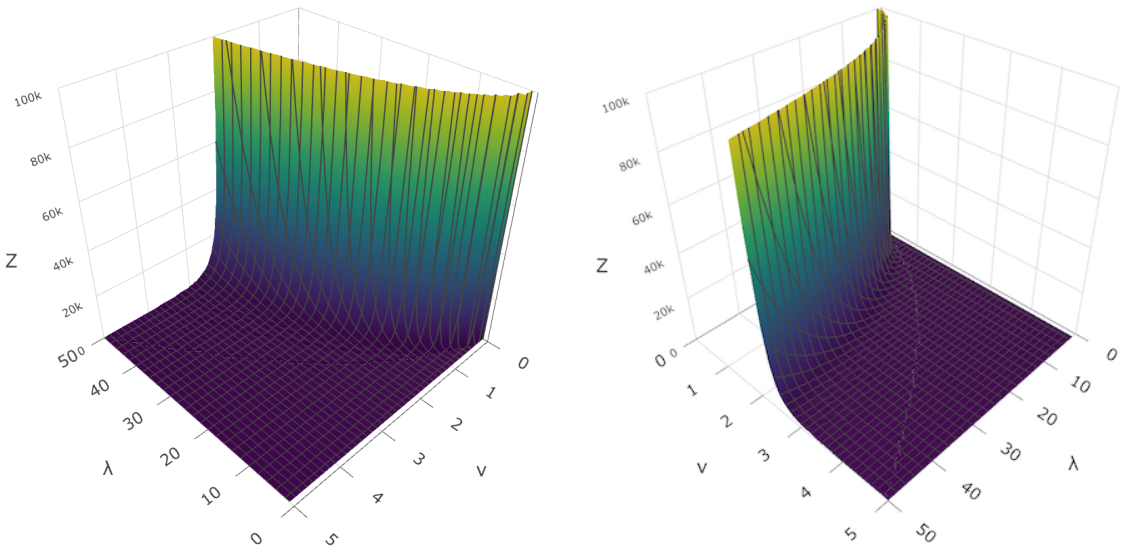


Figure 4.1: Surface plot of the normalizing constant $\mathcal{Z}(\lambda, \nu)$ in the Conway-Maxwell-Poisson distribution with respect to the rate parameter λ and the dispersion parameter ν. (a) and (b) are from different angles of view. The values of $\mathcal{Z}$ are cut off at 10^5.

to display meaningful details. The calculation of $\mathcal{Z}$ values was based on the formula $\mathcal{Z}(\lambda,\nu)=\sum_{x=0}^{\infty}\frac{\lambda^x}{(x!)^\nu}$, where the maximum value of x was taken as 150 for overall convergence of $\mathcal{Z}(\lambda,\nu)$. The normalizing constant $\mathcal{Z}(\lambda,\nu)$ generally converges, but there are parameter combinations (λ,ν) that lead to $\mathcal{Z}$ values converged to extremely large values, particularly in cases of over-dispersion ($\nu<1$). Notably, when λ is relatively large and ν approaches near 0, some $\mathcal{Z}$ values can reach the magnitude of 10^{100}. This somewhat corresponds to the comment in Shmueli et al. (2005) that "When $\nu=0$ and $\lambda\geq 1$, $\mathcal{Z}(\lambda,\nu)$ does not converge, and hence the distribution is undefined." In situations of under-dispersion ($\nu>1$), the $\mathcal{Z}$ values generally fall within the range of less than or around the magnitude of 10^{10}. Consequently, the Conway-Maxwell-Poisson distribution may degenerate if the parameters exceed certain value ranges (Sur et al. 2015). This is also the reason that the dispersion estimates were bounded within the model computations in both Chapter 2 and Chapter 3, an aspect often overlooked in the Conway-Maxwell-Poisson modeling.

In this dissertation, a genuine compliment should be attributed to Huang (2017) for introducing the mean-parameterized Conway-Maxwell-Poisson distribution, centered around the true mean. Without the contribution of Huang (2017), this dissertation might not have been possible, and particularly, it would have been challenging to reveal the advantage of mean-parameterized Conway-Maxwell-Poisson mixtures in handling under-dispersion. Unlike the models involving only a single component, mixture models utilize the EM algorithm for iterative computation, requiring the evaluation of component distributions repeatedly. If a component distribution degenerates due to certain parameter estimates at certain iteration, the EM algorithm may hardly proceed. The establishment of mean-parameterized Conway-Maxwell-Poisson distribution offers an idea to identify the numerical issue encountered in the EM algorithm. The issue was finally solved by bounding the dispersion estimates. The motivation behind bounding the dispersion estimates is that a valid distribution

should always have a valid mean, representing the center of the distribution. As such, degenerated distributions can be rectified by appropriate dispersion estimates. Furthermore, as noted by Sur et al. (2015), some combinations of (λ, ν) may lead to similar distributions. Accordingly, in the Conway-Maxwell-Poisson modeling, multiple sets of parameter estimates might be possible for the same data, an issue known as identifiability. However, this should not be a concern for the mean-parameterized Conway-Maxwell-Poisson distribution, as it is highly unlikely for a single distribution to exhibit multiple distinct centers.

Figure 4.2 presents a surface plot of the mean or expectation, $\mu(\lambda, \nu)$, for the Conway-Maxwell-Poisson distribution, with respect to the original parameters $\lambda \in [1, 50]$ and $\nu \in [0, 5]$. To display more details and the overall trend of the distribution means depending on (λ, ν), the μ values were cut off at 10^2 in 4.2a, 10^4 in 4.2b, and 10^9 in 4.2c. The expectation for the Conway-Maxwell-Poisson distribution shown in Figure 4.2 relates to and behaves consistently with the normalizing constant $\mathcal{Z}(\lambda, \nu)$ shown in Figure 4.1. In the case of over-dispersion ($\nu < 1$), especially when λ is relatively large and ν approaches near 0, some parameter combinations (λ, ν) cause the distribution means extremely large, and even approaching infinity, which is unreasonable for valid mean-parameterized Conway-Maxwell-Poisson distri-

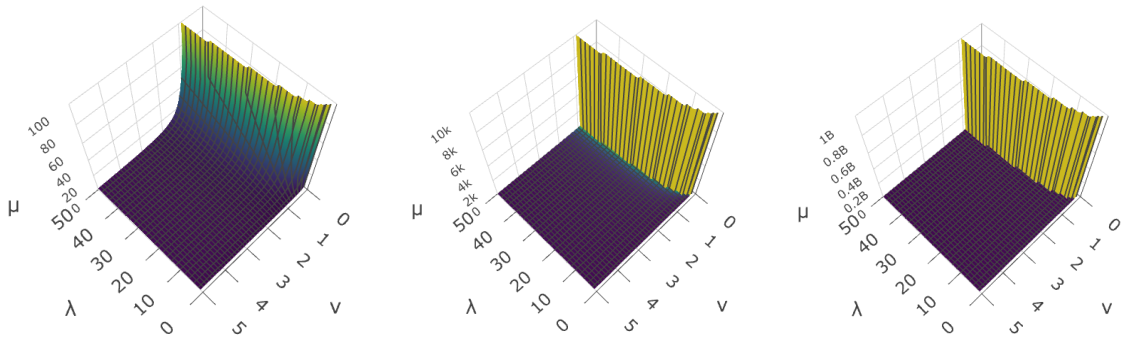


Figure 4.2: Surface plot of the mean or expectation $\mu(\lambda, \nu)$ for the Conway-Maxwell-Poisson distribution with respect to the rate parameter λ and the dispersion parameter ν. The values of μ are cut off at 10^2 in (a), 10^4 in (b), and 10^9 in (c).

butions. For over-dispersion ($\nu < 1$), in order to maintain a reasonable magnitude for mean values, the distribution requires λ relatively small and ν not too small to approach zero. For under-dispersion ($\nu > 1$), the distribution means remain within a reasonable magnitude. This might explain the findings in this dissertation, demonstrating the advantage of mixtures of mean-parameterized Conway-Maxwell-Poisson distributions over Poisson mixtures and negative binomial mixtures in addressing under-dispersion, while showing less capability in addressing over-dispersion compared to negative binomial mixtures. For future applications of mean-parameterized Conway-Maxwell-Poisson distribution, it might be necessary to establish a boundary space with more convincing details for the parameters (λ, ν) or (μ, ν) to ensure the validity of the distribution.

4.2.2 Implication and Improvements

The univariate mixture setting in Chapter 2 demonstrated that the mean-parameterized Conway-Maxwell-Poisson mixture model competently accounted for dispersions over Poisson mixtures. It clearly outperformed negative binomial mixtures for under-dispersed data. However, unlike the univariate setting, the mixtures of mean-parameterized Conway-Maxwell-Poisson regressions in Chapter 3 did not show as notable an advantage over mixtures of Poisson regressions and mixtures of negative binomial regressions. In the case of regression setting, whether the data was over-dispersed or under-dispersed, mixtures of mean-parameterized Conway-Maxwell-Poisson regressions generally showed only slightly better likelihood values than the other two regression models. However, the mixture of Poisson regressions was often the best choice because of a lesser penalty on the Bayesian information criterion (BIC). It's worth noting that mixtures of negative binomial regressions may not be more impressive than mixtures of Poisson regressions for over-dispersion. Nevertheless, we turned to the differences in BIC values for

comparing the three mixture of regressions models.

An idea is postulated to explain why mixtures of mean-parameterized Conway-Maxwell-Poisson (regressions) models performed well in the univariate setting, but less adequately in the regression setting. The dispersion parameter of the mean-parameterized Conway-Maxwell-Poisson distribution largely determines the shape of the distribution. Intuitively, it is challenging to perceive the overall shape of a distribution using information from only a single data point. In the regression setting, where each individual response is assumed to follow an individual mean-parameterized Conway-Maxwell-Poisson distribution, the dispersion parameter may not be as crucial to model, as there is only a single data point involved in the assumed distribution. The univariate setting worked impressively because it involves many responses to determine a single mean-parameterized Conway-Maxwell-Poisson distribution for a component, and thus the dispersion parameter can better model the shape of the distribution. A similar postulation may apply to the comparison between mixtures of negative binomial (regressions) models and mixtures of Poisson (regressions) models regarding the discrepancy in univariate setting and regression setting. Nonetheless, this postulation is challenging to prove. Approaching this issue in a productive manner may provide further insights and directions for improvement.

In Chapter 3, the estimates for the regression setting were obtained using the EM algorithm incorporated with the `nloptr` package (Johnson, 2014), which is an R interface to a number of nonlinear optimization routines. The advantage of using `nloptr` along with the algorithm `NLOPT_LD_SLSQP` is that it allows for constraint equations and solution boundaries, making it suitable for solving the series of nonlinear functions in Chapter 3. However, it is important to note that there is no guarantee that one optimization routine is superior to all others. Moreover, given that the rate parameters λs and the dispersion parameters νs together contribute to valid distributions, the bounds for estimates, especially for dispersion estimates, employed in

Chapter 3 may not cover all potential estimates or the best estimates for the model, although the solutions in Chapter 3 never encountered numerical issues, nor did they produce any error or warning messages.

In a similar manner to solving the mixture of negative binomial regressions model in Chapter 3, a new strategy for solving the mixture of mean-parameterized Conway-Maxwell-Poisson regressions model was developed using the `glmmTMB()` function within the `glmmTMB` package (Brooks et al. 2017). It is fortunate that the `glmmTMB` package was updated in April 2023, allowing for generalized linear regression to the true mean of a single Conway-Maxwell-Poisson distribution. Specifically, the posterior membership probabilities (z_{ij}) are used as weights within `glmmTMB()`, and the optimized solution for each $j = 1, \ldots, m$ component yields the estimates for the $\boldsymbol{\beta}_j$s and ν_js in equations 3.14 and 3.15. The work in Chapter 3 was improved accordingly, primarily by allowing a more lenient range of dispersion estimates. Further details are included in the article by Zhan and Young (2023b). The mixture of mean-parameterized Conway-Maxwell-Poisson regressions model, utilizing the `glmmTMB()` function, was found to be competitive for modeling SIDS data, as illustrated in the next section.

4.2.3 SIDS Data

Symons et al. (1983) presented an analysis on the spatial occurrence of sudden infant death syndrome (SIDS) across North Carolina counties over a four-year period. A Poisson mixture model was used to reveal the epidemiologic information as normal or high-risk for SIDS among different regions in the area. The numbers of live births and SIDS deaths were recorded for $n = 100$ counties in North Carolina from July 1, 1974 to June 30, 1978. Each county has an observed data point. Since SIDS cases are rare incidences, the number of SIDS deaths can be regarded as a count variable, while the corresponding number of live births can be treated as a covariate. The three

mixtures of regressions models, including mixtures of Poisson regressions, mixtures of negative binomial regressions, and mixtures of MCMP1 regressions, are potential candidates to model the SIDS data. We proceed with analyzing these SIDS data in a manner similar to the analysis of the aphids data in Sect. 3.5.

The SIDS data was modeled using the same three mixtures of count regressions we have been comparing throughout this work. The BIC values for different numbers of components in the models are shown in Table 4.1. In each column with the model category specified, the two-component fit shows as the best with the smallest BIC value. Across the models with the same number of components, the MCMP1 model consistently shows as the best in each row. The two-component mixture of MCMP1 regressions outperforms all of the candidate models, despite having only a slight advantage in terms of the BIC values. For the two-component fits, $\Delta\text{BIC}_{\text{Poisson}} = 0.6942$ and $\Delta\text{BIC}_{\text{NB}} = 1.6467$. Under the ΔBIC rule used in our simulation study in Sect. 3.4, the three two-component mixtures of count regressions are comparable for this application, but the mixture of MCMP1 regressions appears to be favored across the mixture fits with the other number of components. Specifically, the ΔBIC values show bigger differences within $m = 1$ (non-mixture fit) and $m = 3$ mixture models than $m = 2$ mixture models, with the mixture of MCMP1 regressions model as a benchmark.

Table 4.1: BIC values for mixtures of Poisson regressions, mixtures of negative binomial regressions, and mixtures of MCMP1 regressions when those models are fit to the SIDS data

	BIC		
m	Poissons	NBs	MCMP1s
1	637.1011	542.9037	542.2042
2	537.0027	538.0002	**536.3535**
3	550.6334	556.4209	554.7741

The estimates and corresponding estimated standard errors for the SIDS data re-

gressed with two-component mixture of Poisson regressions, two-component mixture of negative binomial regressions, and two-component mixture of MCMP1 regressions are reported in Table 4.2. The three mixture models yield similar mixing proportion estimates: $(0.6251, 0.3749)$ for the Poisson components, $(0.5405, 0.4595)$ for the negative binomial components, and $(0.5184, 0.4816)$ for the MCMP1 components. Unlike the estimates in the aphids data analysis where the estimated standard errors for the mixing proportion estimates are of similar magnitude across the models, the mixing proportion estimates in the SIDS data analysis have relatively larger variation for both the mixture of negative binomial regressions and the mixture of MCMP1 regressions when compared to the mixture of Poisson regressions. For the mixture of Poisson regressions, the estimated standard error for the mixing proportions is as small as 0.0098. For the mixture of negative binomial regressions and the mixture of MCMP1 regressions, the estimated standard errors for the mixing proportions are 0.1203 and 0.1655, respectively. The regression estimates along with the estimated standard errors are slightly different, but similar in magnitude across the models. The mixture of negative binomial regressions yields one extremely large dispersion

Table 4.2: The estimates and corresponding estimated standard errors for the SIDS data fit using the two-component mixture of Poisson regressions, mixture of negative binomial (NB) regressions, and mixture of MCMP1 regressions

	Poissons		NBs		MCMP1s	
Par.	Estimate	$\widehat{\text{SE}}$	Estimate	$\widehat{\text{SE}}$	Estimate	$\widehat{\text{SE}}$
π_1	0.6251	0.0098	0.5405	0.1203	0.5184	0.1655
π_2	0.3749	0.0098	0.4595	0.1203	0.4816	0.1655
β_{01}	0.0021	0.2466	-0.0768	0.3907	-0.0444	0.4053
β_{11}	4.42×10^{-4}	7.52×10^{-5}	4.58×10^{-4}	1.52×10^{-4}	4.53×10^{-4}	1.07×10^{-4}
β_{02}	1.8556	0.2261	1.6982	0.3051	1.6593	0.4156
β_{12}	8.64×10^{-5}	1.85×10^{-5}	9.76×10^{-5}	2.23×10^{-5}	9.85×10^{-5}	4.66×10^{-5}
ν_1	—	—	37297.5	23651.3	1.2994	16.2954
ν_2	—	—	9.3982	82394.8	0.5744	7.2049
$\ell_o^{(\infty)}$	-256.9884		-252.8820		-252.0587	

estimate, which implies an approximated Poisson component being estimated. The mixture of MCMP1 regressions produces reasonable estimates for the component dispersions, which implies the first component is under-dispersed with estimate 1.2994 (greater than 1) and the second component is over-dispersed with estimate 0.5744 (less than 1). If we again consider removing the top 2% of the bootstrap samples when they are sorted according to their dispersion estimates (from largest to smallest), then the estimated standard errors for the dispersion parameters are $\widehat{\text{SE}}(\hat{\nu}_1) = 1.0675$ and $\widehat{\text{SE}}(\hat{\nu}_2) = 2.8868$, both of which are noticeably lower than the estimated standard errors reported in Table 4.2. Similar to the aphids data in Table 3.6, the loglikelihood values across the models in the row are close for the SIDS data in Table 4.2, but the mixture of MCMP1 regressions has the largest loglikelihood value, while the mixture of Poisson regressions has the smallest loglikelihood value.

Figure 4.3 is a scatterplot of the SIDS data overlaid with the estimated two-component mixture of MCMP1 regressions model, which is the best fit according to Table 4.1. The other mixture model fits are not included on this scatterplot because their regression mean lines are just slightly different from that of the two-component

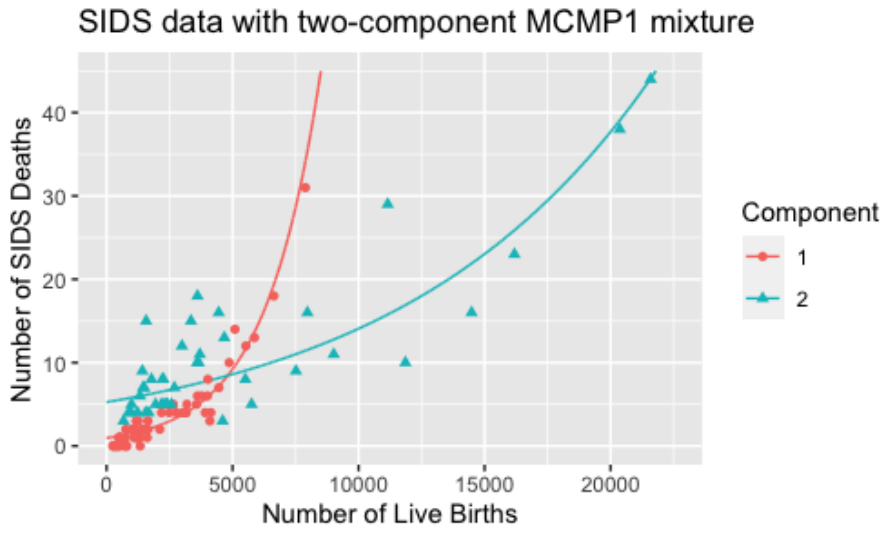


Figure 4.3: Scatterplot of the SIDS data overlaid with the conditional mean lines estimated for the two-component mixture of MCMP1 regressions model

mixture of MCMP1 regressions; see Table 4.2 where the regression estimates across the models are only slightly different. From Figure 4.3, the $n = 100$ counties in North Carolina are clearly differentiated by two components. The red component indicates the high risk to SIDS conditioned on the live births for one group of counties. The identified high-risk counties have the data points under-dispersed based on the dispersion estimate in Table 4.2. The under-dispersion sometimes implies some abnormality, regarding SIDS in the group of counties. The blue component indicates the risk to SIDS more typically normal within the other group of counties, considering the moderate-ascending SIDS deaths conditioned on the live births displayed by the component regression line. This group of counties has the data points over-dispersed, which is commonly observed with count data. The mixture of MCMP1 regressions is not just effective for cluster analysis, but also helpful to reveal the dispersion information about the components.

4.3 Future Research

4.3.1 Regressions Extended on Mixing Proportions

Chapter 3 in this dissertation introduces a mixture of regressions model within the framework of generalized linear models. This regression model assumes that each response is generated from a mean-parameterized Conway-Maxwell-Poisson distribution, with its mean linked to the respective covariates through a log-linear relationship. In finite mixture models, the membership of an observation to a component can be treated as a categorical variable, making logistic regression a feasible choice for modeling.

Future work may extend the model presented in Chapter 3 to a more versatile regression setting by allowing the mixing proportions to be modeled on the related predictors. This extension is novel and may hold practical application, especially considering the mixing proportions potentially depending on some concomitant vari-

ables, and the limited existing research in this area (Young and Hunter 2010; Huang and Yao 2012). Accordingly, we propose a new regression model, briefly introduced here for future study.

The new m-component mixture of mean-parameterized Conway-Maxwell-Poisson regressions model is constructed based on the conditional distribution of $Y|(\boldsymbol{X}, \tilde{\boldsymbol{X}})$. Here, $Y \in \mathbb{N}$ is the discrete response variable, $\boldsymbol{X} = (X_1, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$ is a p-dimensional covariate vector, and $\tilde{\boldsymbol{X}} = (\tilde{X}_1, \ldots, \tilde{X}_q)^{\mathrm{T}} \in \mathbb{R}^q$ is a q-dimensional concomitant vector.

The mean parameters for the components, denoted as μ_j for $j = 1, \ldots, m$, are modeled as a function of the covariates via a log link function

$$\mu_j = \exp(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_j), \tag{4.1}$$

where $\boldsymbol{x} = (x_0, x_1, \ldots, x_p)^{\mathrm{T}}$ and $\boldsymbol{\beta}_j = (\beta_{0j}, \beta_{1j}, \ldots, \beta_{pj})^{\mathrm{T}}$. Here, $x_0 = 1$ so as to allow for modeling with an intercept. Therefore, the β_{0j} is the intercept for the jth component regression and the $\beta_{1j}, \ldots, \beta_{pj}$ correspond to the coefficients for the respective covariates within the jth component regression.

The mixing proportion parameters, denoted as π_j for $j = 1, \ldots, m$, are modeled as a function of the concomitant variables via a logit function. The component $j = 1$ is taken as the baseline to calculate the odds for component $j = 2, \ldots, m$

$$\log\Big(\frac{\pi_j}{\pi_1}\Big) = \tilde{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{\alpha}_j, \;\; j = 2, \ldots, m, \tag{4.2}$$

which gives

$$\pi_j = \pi_1 \exp(\tilde{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{\alpha}_j), \;\; j = 2, \ldots, m. \tag{4.3}$$

Since $\sum_{j=1}^{m} \pi_j = 1$, then we have

$$\pi_1 = \frac{1}{1 + \sum_{j=2}^{m} \exp(\tilde{\boldsymbol{x}}^{\mathrm{T}} \boldsymbol{\alpha}_j)}, \text{ and } \pi_j = \frac{\exp(\tilde{\boldsymbol{x}}^{\mathrm{T}} \boldsymbol{\alpha}_j)}{1 + \sum_{j=2}^{m} \exp(\tilde{\boldsymbol{x}}^{\mathrm{T}} \boldsymbol{\alpha}_j)} \text{ for } j = 2, \ldots, m, \quad (4.4)$$

where $\tilde{\boldsymbol{x}} = (\tilde{x}_0, \tilde{x}_1, \ldots, \tilde{x}_q)^{\mathrm{T}}$ and $\boldsymbol{\alpha}_j = (\alpha_{0j}, \alpha_{1j}, \ldots, \alpha_{qj})^{\mathrm{T}}$. Similarly, $\tilde{x}_0 = 1$ is for modeling an intercept. The α_{0j} is the intercept of the jth component regression and the $\alpha_{1j}, \ldots, \alpha_{qj}$ correspond to the coefficients for the respective concomitant variables within the jth component regression.

Recall that the mixture density of mean-parameterized Conway-Maxwell-Poisson distributions is

$$g(y; \boldsymbol{x}, \tilde{\boldsymbol{x}}, \boldsymbol{\Psi}) = \sum_{j=1}^{m} \pi_j \frac{\lambda(\mu_j, \nu_j)^y}{(y!)^{\nu_j}} \frac{1}{\mathcal{Z}(\lambda(\mu_j, \nu_j), \nu_j)}. \quad (4.5)$$

Given the regression settings in 4.1 and 4.4, the m-component mixture of mean-parameterized Conway-Maxwell-Poisson regressions model for $Y|(\boldsymbol{X}, \tilde{\boldsymbol{X}})$ has the mixture density

$$\begin{aligned} g(y; \boldsymbol{x}, \tilde{\boldsymbol{x}}, \boldsymbol{\Psi}) = & \frac{1}{1 + \sum_{j=2}^{m} \exp(\tilde{\boldsymbol{x}}^{\mathrm{T}} \boldsymbol{\alpha}_j)} \frac{\lambda(\exp(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}_1), \nu_1)^y}{(y!)^{\nu_1}} \frac{1}{\mathcal{Z}(\lambda(\exp(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}_1), \nu_1), \nu_1)} + \\ & \sum_{j=2}^{m} \frac{\exp(\tilde{\boldsymbol{x}}^{\mathrm{T}} \boldsymbol{\alpha}_j)}{1 + \sum_{j=2}^{m} \exp(\tilde{\boldsymbol{x}}^{\mathrm{T}} \boldsymbol{\alpha}_j)} \frac{\lambda(\exp(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}_j), \nu_j)^y}{(y!)^{\nu_j}} \frac{1}{\mathcal{Z}(\lambda(\exp(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}_j), \nu_j), \nu_j)}, \end{aligned} \quad (4.6)$$

where the parameter vector is

$$\boldsymbol{\Psi} = \left(\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_m^{\mathrm{T}}, \boldsymbol{\alpha}_1^{\mathrm{T}}, \ldots, \boldsymbol{\alpha}_m^{\mathrm{T}}, \nu_1, \ldots, \nu_m\right)^{\mathrm{T}}.$$

Here, the ν_js are the dispersion parameters with assumption that the data points in the same component follow this degree of dispersion.

4.3.2 Singularity Considerations

The singularity issue was not a concern in my study, because Fisher information matrices were not involved in the computation work. In this dissertation, the real data were fitted with multiple-component mixtures for choosing the appropriate number of components in the mixture models. The results demonstrated the effectiveness of BIC in selecting a suitable model from different numbers of components (see Table 2.3, Table 3.5, and Table 4.1). Moreover, my unpublished results from this study indicated that having more components than necessary didn't lead to strange estimates. The extra component may have a mixing proportion estimate of less than 0.05, but the other estimates for the extra component were reasonable. The loglikelihood values for models with extra components were almost the same as that from the model with appropriate number of components. In other words, the loglikelihood converged as more components were added to the mixture, but the models with more components were not selected due to the BIC penalty resulting from the inclusion of more parameters. In the univariate setting in Chapter 2, models with additional components produced fit plots, which almost overlaid with that from the best model having the appropriate number of components.

Singularity is often identified in certain statistical models when the Fisher information matrix becomes singular. As a consequence of singularity, parameter estimates may exhibit unusual behaviors since the Cramér-Rao theorem doesn't hold well, thus affecting subsequent procedures for hypothesis testing, model selection, and inference. Whether singularity exists and how it behaves is a case-by-case and model-by-model matter, particularly for mixture models. In the case of Gaussian mixtures, fitting a Gaussian component to only one data point can lead to a variance estimate of zero, causing a singular Fisher information matrix. Accordingly, the likelihood for the component goes toward infinity, and well-posed solutions may not be guaranteed under the framework of maximum likelihood estimation.

Singular BIC, a generalization of BIC and a method to address the singularity issue in mixture models, was studied by Drton and Plummer (2017) and discussed with some other researchers in the same manuscript, which was available in 2013. The corresponding theory behind this approach, known as the singular learning theory, was proposed earlier by Watanabe (2009), who is established by his extensive research on model selection criteria. In the presence of singularity, maximum likelihood estimator does not behave reasonably, even though the maximum likelihood method itself is still acceptable, but might yield unusual estimates. The singular BIC, incorporating loglikelihood in its formula, is one application of the singular learning theory. This theory can be applied to various models with singularity issues, not limited to mixture models.

While likelihood solutions are generally consistent for mixture models estimated using the EM algorithm (Redner and Walker 1984), and ordinary BIC is consistent as well (Keribin 2000), likelihood estimation may not perform well when the loglikelihood appears unusual under singular conditions. The concept of singular BIC can assist in obtaining accurate likelihood calculations for mixture models when singularity occurs. The first step to apply singular BIC is to determine whether the model is singular or not. It yields the normal BIC if the model is not singular. Singular BIC is useful only under singular conditions, and it can be challenging to estimate the learning coefficient and multiplicity number for each model. Examples provided by Drton and Plummer (2017) suggest that these values are often approximated reasonably, and that somehow depends on the prior or data-generating distributions, and especially the number of components in mixture models. Singular BIC makes sense when dealing with singular issues, although it may be challenging to apply due to the need to identify learning coefficient and multiplicity number on a model-by-model basis.

In this dissertation, my study focused on heterogeneous data that visibly consists

of multiple components. However, in some applications of mixture models, it's not easy to distinguish between homogeneity (one component) or distinct components in the population. Accordingly, a procedure to test for homogeneity or to determine the exact number of components is required. The likelihood-ratio test (LRT) is a common statistical tool for model selection, but encounters difficulties in mixture models, concerning issues such as limiting distribution, identifiability and others (Dacunha-Castelle and Gassiat 1999). The asymptotic behaviour of the LRT statistics for two-component mixtures was studied, and a bootstrap procedure was used to obtain p-value of the LRT (Chen and Chen 2001). Certainly, the concept of singular BIC, with its improved likelihood calculation, may help the LRT procedure in choosing the number of components for mixture models when singularity is involved.

Bibliography

M. A. Abdel-Aty and A. E. Radwan. Modeling traffic accident occurrence and involvement. *Accident Analysis & Prevention*, 32(5):633–642, 2000.

M. Aitken. A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, 6(3):251–262, 1996.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, pages 267–281. Akademiai Kiado, Budapest, Hungary, 1973.

A. Anandkumar, D. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden Markov models. In *25th Annual Conference on Learning Theory*, volume 23:33, pages 1–34, 2012.

M. Arora, N. R. Chaganty, and K. F. Sellers. A flexible regression model for zero- and k-inflated count data. *Journal of Statistical Computation and Simulation*, 91 (9):1815–1845, 2021.

M. F. Bachmann, G. Köhler, B. Ecabert, T. W. Mak, and M. Kopf. Cutting edge: Lymphoproliferative disease in the absence of CTLA-4 is not T cell autonomous. *Journal of Immunology*, 163(3):1128–1131, 1999.

G. E. Bardwell and E. L. Crow. A two-parameter family of hyper-Poisson distributions. *Journal of the American Statistical Association*, 59(305):133–141, 1964.

T. Benaglia, D. Chauveau, D. R. Hunter, and D. S. Young. **mixtools**: An **R** package for analyzing mixture models. *Journal of Statistical Software*, 32(6):1–29, 2010.

C. Biernacki, G. Celeux, and G. Govaert. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3-4):561–575, 2003.

P. Boatwright, S. Borle, and J. B. Kadane. A model of the joint distribution of purchase quantity and timing. *Journal of the American Statistical Association*, 98 (463):564–572, 2003.

R. Brey and J. L. Walker. Latent temporal preferences: An application to airline travel. *Transportation Research Part A: Policy and Practice*, 45(9):880–895, 2011.

M. E. Brooks, K. Kristensen, K. J. van Benthem, A. Magnusson, C. W. Berg, A. Nielsen, H. J. Skaug, M. Mächler, and B. M. Bolker. `glmmTMB` balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *R Journal*, 9(2):378–400, 2017. URL `https://doi.org/10.32614/RJ-2017-066`.

J. Castillo and M. Pérez-Casany. Weighted Poisson distributions for overdispersion and underdispersion situations. *Annals of the Institute of Statistical Mathematics*, 50:567–585, 1998.

G. Celeux and G. Soromenho. An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13(2):195–212, 1996.

H. Chen and J. Chen. The likelihood ratio test for homogeneity in finite mixture models. *Canadian Journal of Statistics*, 29(2):201–215, 2001.

H. Chen, J. Chen, and J. D. Kalbfleisch. A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 63(1):19–29, 2001.

P. C. Consul. *Generalized Poisson Distributions: Properties and Applications*. Marcel Dekker, Inc., New York, NY, 1989.

P. C. Consul and G. C. Jain. A generalization of the Poisson distribution. *Technometrics*, 15(4):791–799, 1973.

R. W. Conway and W. L. Maxwell. A queuing model with state dependent service rates. *Journal of Industrial Engineering*, 12:132, 1961.

D. R. Cox. Some remarks on overdispersion. *Biometrika*, 70(1):269–274, 1983.

I. J. Cox, J. Ghosn, and P. N. Yianilos. Feature-based face recognition using mixture-distance. In *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 209–216. IEEE, 1996.

R. B. Cunningham and D. B. Lindenmayer. Modeling count data of rare species: Some statistical issues. *Ecology*, 86(5):1135–1142, 2005.

D. Dacunha-Castelle and E. Gassiat. Testing the order of a model using locally conic parametrization: Population mixtures and stationary ARMA processes. *Annals of Statistics*, 27(4):1178–1209, 1999.

D. G. Daniel, T. E. Goldberg, R. D. Gibbons, and D. R. Weinberger. Lack of a bimodal distribution of ventricular size in schizophrenia: A Gaussian mixture analysis of 1056 cases and controls. *Biological Psychiatry*, 30(9):887–903, 1991.

R. D. De Veaux. Mixtures of linear regressions. *Computational Statistics & Data Analysis*, 8(3):227–245, 1989.

C. B. Dean and E. R. Lundy. Overdispersion. In *Wiley StatsRef: Statistics Reference Online*, 2016.

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

F. V. Dénes, L. F. Silveira, and S. R. Beissinger. Estimating abundance of unmarked animal populations: Accounting for imperfect detection and other sources of zero inflation. *Methods in Ecology and Evolution*, 6(5):543–556, 2015.

M. Drton and M. Plummer. A Bayesian information criterion for singular models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(2): 323–380, 2017.

T. W. Epps and M. L. Epps. The stochastic dependence of security price changes and transaction volumes: Implications for the mixture-of-distributions hypothesis. *Econometrica*, 44(2):305–321, 1976.

W. Feng, Y. Liu, J. Wu, K. P. Nephew, T. H. Huang, and L. Li. A Poisson mixture model to identify changes in RNA polymerase II binding quantity using high-throughput sequencing technology. *BMC Genomics*, 9(Suppl 2):S23, 2008.

C. Fraley and A. E. Raftery. How many clusters? Which clustering method to use? Answers via model-based cluster analysis. *Computer Journal*, 41(8):578–588, 1998.

S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Sringer, New York, NY, 2006.

D. Gerogiannis, C. Nikou, and A. Likas. The mixtures of student's t-distributions as a robust framework for rigid registration. *Image and Vision Computing*, 27(9): 1285–1294, 2009.

S. M. Goldfeld and R. E. Quandt. A Markov model for switching regressions. *Journal of Econometrics*, 1(1):3–15, 1973.

L. Grilli, C. Rampichini, and R. Varriale. Binomial mixture modeling of university credits. *Communications in Statistics – Theory and Methods*, 44(22):4866–4879, 2015.

B. Grün and F. Leisch. Finite mixtures of generalized linear regression models. In *Recent Advances in Linear Models and Related Areas: Essays in Honour of Helge Toutenburg*, pages 205–230, Heidelberg, 2008. Physica-Verlag HD.

S. D. Guikema and J. P. Coffelt. A flexible count data regression model for risk analysis. *Risk Analysis: An International Journal*, 28(1):213–223, 2008.

V. Hasselblad. Estimation of parameters for a mixture of normal distributions. *Technometrics*, 8(3):431–444, 1966.

J. M. Hilbe. *Negative Binomial Regression*. Cambridge University Press, Cambridge, UK, 2nd edition, 2011.

J. Hinde and C. G. Demétrio. Overdispersion: Models and estimation. *Computational Statistics & Data Analysis*, 27(2):151–170, 1998.

A. Huang. Mean-parametrized Conway-Maxwell-Poisson regression models for dispersed counts. *Statistical Modelling*, 17(6):359–380, 2017.

A. Huang and P. J. Rathouz. Orthogonality of the mean and error distribution in generalized linear models. *Communications in Statistics – Theory and Methods*, 46 (7):3290–3296, 2017.

C. Huang, X. Liu, T. Yao, and X. Wang. An efficient EM algorithm for the mixture of negative binomial models. *Journal of Physics: Conference Series*, 1324(1):012093, 2019.

M. Huang and W. Yao. Mixture of regression models with varying mixing proportions: A semiparametric approach. *Journal of the American Statistical Association*, 107 (498):711–724, 2012.

M. Hurn, A. Justel, and C. P. Robert. Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics*, 12(1):55–79, 2003.

N. Ismail, K. M. Mohd Ali, and A. C. Chiew. A model for insurance claim count with single and finite mixture distribution. *Sains Malaysiana*, 33(2):173–194, 2004.

R. Jiang and D. N. P. Murthy. Modeling failure-data by mixture of 2 weibull distributions: A graphical approach. *IEEE Transactions on Reliability*, 44(3):477–488, 1995.

S. G. Johnson. The **NLopt** nonlinear-optimization package, 2014. URL `https://github.com/stevengj/nlopt`.

H. Kasahara and K. Shimotsu. Testing the number of components in normal mixture regression models. *Journal of the American Statistical Association*, 110(512):1632–1645, 2015.

B. C. Kelly, M. Vestergaard, and X. Fan. Determining quasar black hole mass functions from their broad emission lines: Application to the bright quasar survey. *Astrophysical Journal*, 692(2):1388, 2009.

C. Keribin. Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 49–66, 2000.

S. J. Kon. Models of stock returns – a comparison. *Journal of Finance*, 39(1):147–165, 1984.

F. Leisch. **FlexMix**: A general framework for finite mixture models and latent class regression in **R**. *Journal of Statistical Software*, 11(8):1–18, 2004.

B. G. Leroux. Consistent estimation of a mixing distribution. *Annals of Statistics*, 20(3):1350–1360, 1992.

T. W. Lewis, B. M. Wiles, A. M. Llewellyn-Zaidi, K. M. Evans, and D. G. O'Neill. Longevity and mortality in Kennel Club registered dog breeds in the UK in 2014. *Canine Genetics and Epidemiology*, 5(1):10, 2018.

J. Li and H. Zha. Two-way Poisson mixture models for simultaneous document classification and word clustering. *Computational Statistics & Data Analysis*, 50 (1):163–180, 2006.

Q. Li, J. R. Noel-MacDonnell, D. C. Koestler, E. L. Goode, and B. L. Fridley. Subject level clustering using a negative binomial model for small transcriptomic studies. *BMC Bioinformatics*, 19(1):474, 2018.

X. Li and D. K. Dey. Estimation of COVID-19 mortality in the United States using Spatio-temporal Conway-Maxwell-Poisson model. *Spatial Statistics*, 49(100542): 1–11, 2022.

B. G. Lindsay. *Mixture Models: Theory, Geometry and Applications*, volume 5 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics and the American Statistical Association, 1995.

D. Lord, S. D. Guikema, and S. R. Geedipally. Application of the Conway–Maxwell–Poisson generalized linear model for analyzing motor vehicle crashes. *Accident Analysis & Prevention*, 40(3):1123–1134, 2008.

H. J. Lynch, J. T. Thorson, and A. O. Shelton. Dealing with under-and over-dispersed count data in life history, spatial, and community ecology. *Ecology*, 95(11):3173–3180, 2014.

G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, 2nd edition, 2007.

G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley-Interscience, New York, 2000.

G. J. McLachlan, S. X. Lee, and S. I. Rathnayake. Finite mixture models. *Annual Review of Statistics and Its Application*, 6(1):355–378, 2019.

P. D. McNicholas. *Mixture Model-Based Classification*. Taylor & Francis, UK, 2016.

V. Melnykov and R. Maitra. Finite mixture models and model-based clustering. *Statistics Surveys*, 4:80–116, 2010.

D. G. Muenz, T. M. Braun, and J. M. Taylor. Modeling adverse event counts in phase I clinical trials of a cytotoxic agent. *Clinical Trials*, 15(4):386–397, 2018.

P. Papastamoulis, M. L. Martin-Magniette, and C. Maugis-Rabusseau. On the estimation of mixtures of Poisson regression models with large number of components. *Computational Statistics & Data Analysis*, 93:97–106, 2016.

B. J. Park and D. Lord. Application of finite mixture models for vehicle crash data analysis. *Accident Analysis & Prevention*, 41(4):683–691, 2009.

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

L. S. C. Piancastelli, N. Friel, W. Barretto-Souza, and H. Ombao. Multivariate Conway-Maxwell-Poisson distribution: Sarmanov method and doubly intractable Bayesian inference. *Journal of Computational and Graphical Statistics*, 32(2):483–500, 2022.

R. E. Quandt. A new approach to estimating switching regressions. *Journal of the American Statistical Association*, 67(338):306–310, 1972.

A. E. Raftery. Bayesian model selection in social research. *Sociological Methodology*, 25:111–163, 1995.

R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SSIAM review*, 26(2):195–239, 1984.

E. E. Ribeiro Jr, W. M. Zeviani, W. H. Bonat, C. G. Demetrio, and J. Hinde. Reparametrization of COM-Poisson regression models with applications in the analysis of experimental data. *Statistical Modelling*, 20(5):443–466, 2020.

J. Rodrigues, M. de Castro, V. G. Cancho, and N. Balakrishnan. COM-Poisson cure rate survival models and an application to a cutaneous melanoma data. *Journal of Statistical Planning and Inference*, 139(10):3605–3611, 2009.

SAS Institute Inc. *SAS/ETS 13.1 User's Guide*. SAS Publishing, 2013.

P. Schlattmann. *Medical Applications of Finite Mixture Models*. Statistics for Biology and Health. Springer Berlin, Heidelberg, 2009.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

K. F. Sellers. *The Conway–Maxwell–Poisson Distribution*. Cambridge University Press, Cambridge, UK, 2023.

K. F. Sellers and A. Raim. A flexible zero-inflated model to address data dispersion. *Computational Statistics & Data Analysis*, 99:68–80, 2016.

K. F. Sellers and G. Shmueli. A flexible regression model for count data. *Annals of Applied Statistics*, 4(2):943–961, 2010.

K. F. Sellers and G. Shmueli. Data dispersion: Now you see it... now you don't. *Communications in Statistics – Theory and Methods*, 42(17):3134–3147, 2013.

K. F. Sellers and D. S. Young. Zero-inflated sum of Conway-Maxwell-Poissons (ZIS-CMP) regression. *Journal of Statistical Computation and Simulation*, 89(9):1649–1673, 2019.

G. Shmueli, T. P. Minka, J. B. Kadane, S. Borle, and P. Boatwright. A useful distribution for fitting discrete data: Revival of the Conway–Maxwell–Poisson distribution.

Journal of the Royal Statistical Society Series C: Applied Statistics, 54(1):127–142, 2005.

S. Shoham, M. R. Fellows, and R. A. Normann. Robust, automatic spike sorting using mixtures of multivariate t-distributions. *Journal of Neuroscience Methods*, 127(2):111–122, 2003.

G. K. Smyth and B. Jørgensen. Fitting Tweedie's compound Poisson model to insurance claims data: Dispersion modelling. *ASTIN Bulletin: The Journal of the IAA*, 32(1):143–157, 2002.

M. Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

P. Sur, G. Shmueli, S. Bose, and P. Dubey. Modeling bimodal discrete data using Conway-Maxwell-Poisson mixture models. *Journal of Business & Economic Statistics*, 33(3):352–365, 2015.

M. J. Symons, R. C. Grimson, and Y. C. Yuan. Clustering of rare events. *Biometrics*, 39(1):193–205, 1983.

H. Teicher. Identifiability of finite mixtures. *Annals of Mathematical Statistics*, pages 1265–1269, 1963.

T. R. Turner. Estimating the propagation rate of a viral infection of potato plants via mixtures of regressions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(3):371–384, 2000.

H. K. Ünlü, D. S. Young, A. Yiğiter, and L. H. Özcebe. A mixture model with Poisson and zero-truncated Poisson components to analyze road traffic accidents in Turkey. *Journal of Applied Statistics*, 49(4):1003–1017, 2022.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with **S***. Springer, New York, 4th edition, 2002.

K. Viele and B. Tong. Modeling with mixtures of linear regressions. *Statistics and Computing*, 12(4):315–330, 2002.

S. Watanabe. *Algebraic Geometry and Statistical Learning Theory*, volume 25. Cambridge university press, 2009.

W. Whitt. On approximations for queues, III: Mixtures of exponential distributions. *AT&T Bell Laboratories Technical Journal*, 63(1):163–175, 1984.

C. F. J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11(1):95–103, 1983.

Y. Wu and P. Yang. Optimal estimation of Gaussian mixtures via denoised method of moments. *arXiv e-prints*, arXiv:1807.07237 [math.ST], 2018.

S. J. Yakowitz and J. D. Spragins. On the identifiability of finite mixtures. *Annals of Mathematical Statistics*, 39(1):209–214, 1968.

K. C. H. Yip and K. K. W. Yau. On modeling claim frequency data in general insurance with extra zeros. *Insurance: Mathematics and Economics*, 36(2):153–163, 2005.

D. S. Young and D. R. Hunter. Mixtures of regressions with predictor-dependent mixing proportions. *Computational Statistics & Data Analysis*, 54(10):2253–2266, 2010.

D. Zhan and D. S. Young. Finite mixtures of mean-parameterized Conway-Maxwell-Poisson models. *Statistical Papers*, in press, 2023a.

D. Zhan and D. S. Young. Finite mixtures of mean-parameterized Conway-Maxwell-Poisson regressions. *Journal of Statistical Theory and Practice*, under review, 2023b.

P. Zhang, H. Y. Wu, C. W. Chiang, L. Wang, S. Binkheder, X. Wang, D. Zeng, S. K. Quinney, and L. Li. Translational biomedical informatics and pharmacometrics approaches in the drug interactions research. *CPT: Pharmacometrics & Systems Pharmacology*, 7(2):90–102, 2018.

Y. Zou, Y. Zhang, and D. Lord. Application of finite mixture of negative binomial regression models with varying weight parameters for vehicle crash data analysis. *Accident Analysis & Prevention*, 50:1042–1051, 2013.

Vita

Dongying Zhan

Education:

- University of Kentucky, Lexington, KY
  Ph.D. in Statistics, expected December 2023
- University of Kentucky, Lexington, KY
  M.S. in Statistics, May 2020
- Institute of Mechanics, Chinese Academy of Sciences, Beijing, China
  Ph.D. in Biomechanics, June 2009

Professional Positions:

- Teaching Assistant, Department of Statistics, University of Kentucky, August 2018–December 2023
- Post-doctor, College of Engineering, University of Kentucky, November 2009–June 2011

Publications & Preprints:

- D. Zhan and D. S. Young (2023). "Finite mixtures of mean-parameterized Conway–Maxwell–Poisson regressions." *Journal of Statistical Theory and Practice*, under review.
- D. Zhan and D. S. Young (2023). "Finite mixtures of mean-parameterized Conway–Maxwell–Poisson models." *Statistical Papers*, in press.
- X. Zhang, D. Zhan, and H. Y. Shin (2013). "Integrin subtype-dependent CD18 cleavage under shear and its influence on leukocyte-platelet binding." *Journal of Leukocyte Biology*, 93(2), 251-258.
- D. Zhan, Y. Zhang, and M. Long (2012). "Spreading of human neutrophils on an ICAM-1-immobilized substrate under shear flow." *Chinese Science Bulletin*, 57:769–775.