A contaminated regression model for count health data

Otto, Arnoldus F.[a], Ferreira, Johannes T.[a], Tomarchio, Salvatore D.[c], Bekker, Andriëtte[a,b], Punzo, Antonio[c]

[a]*Department of Statistics, University of Pretoria, Pretoria, South Africa,*
[b]*Centre for Environmental Studies, Department of Geography, Geoinformatics and Meteorology, University of Pretoria, Pretoria, South Africa,*
[c]*Department of Economics and Business, University of Catania, Catania, Italy,*

Abstract

In medical and health research, investigators are often interested in countable quantities such as hospital length of stay (for example in days) or the number of doctor visits. Poisson regression is commonly used to model such count data, but this approach can't accommodate overdispersion — when the variance exceeds the mean. To address this issue, the negative binomial (NB) distribution (NB-D), and by extension, NB regression, provide a well-documented alternative. However, real-data applications present additional challenges that must be considered. Two such challenges are: i) the presence of (mild) outliers that can influence the performance of the NB-D, and ii) the availability of covariates that can enhance inference about the mean of the count variable of interest. To jointly address these issues, we propose the contaminated NB (cNB) distribution that exhibits the necessary flexibility to accommodate mild outliers. This model is shown to be simple yet elegant in application, as well as intuitive in interpretation. In addition to the parameters of the NB-D, our proposed model has a parameter describing the proportion of mild outliers and one specifying the degree of contamination. Then, to allow available covariates to improve the estimation of the mean of the cNB distribution, we propose the cNB regression model. An expectation-maximization algorithm is outlined for parameter estimation, and its performance is evaluated through a parameter recovery study. Additionally, a sensitivity analysis is conducted to investigate the performance of our proposed models. The effectiveness of our model is demonstrated on two health datasets, where it outperforms well-known count models. The methodology proposed in this paper has been implemented in an R package, which is publicly available at https://github.com/arnootto/cNB.

Keywords: kurtosis, mild outliers, negative binomial, overdispersion, skewness

1. Introduction

Count data are routinely encountered across various disciplines including epidemiology, social sciences, and economics [22, 9]; particularly noteworthy is their prevalence in the realm of healthcare and medicine [10, 23, 24]. Investigators are often interested in countable quantities such as hospital length of stay (for example, in days) or the number of doctor visits. Length of stay is often used as an indicator of efficiency, as a shorter stay will reduce the cost per discharge and shift care from inpatient to less expensive post-acute settings [31, 8]. Likewise, the number of doctor visits serves as a measure of healthcare utilization and access, reflecting the frequency and necessity of medical care received by patients [4]. Analyzing these metrics helps assess the overall performance and efficiency of healthcare systems, understand patient behaviour and needs, and identify areas for potential improvement in service delivery and cost management.

The values of the count variables are always nonnegative integers, and the distribution is often skewed. The Poisson regression model (Poisson-RM) is traditionally the first considered method for such data and implies a Poisson assumption for the counts conditional to some covariates. However, it operates under the assumption that the conditional mean and variance of the counts are equal. This drawback limits its applicability in scenarios where overdispersion or "extra-Poisson" variation is evident [6, 15, 19, 3].

As discussed in [14], there are two types of overdispersion: apparent and real. Apparent overdispersion can be remedied by various operations on the data, such as adding appropriate predictor(s), constructing required interactions, and transforming predictor(s) or the response. Conversely, real overdispersion is a problem that affects the reliability of both the model parameter estimates and fit in general. To address the shortcomings of the Poisson distribution (Pois-D) in the presence of real overdispersion, the negative binomial (NB) distribution

(NB-D) has emerged as a viable alternative. The probability mass function (PMF) of the mean parameterized NB-D for a count variable Y is

$$\begin{aligned} f_{\mathsf{NB}}(y;\mu,\alpha) &= \binom{y+1/\alpha-1}{y}\left(\frac{\mu}{\mu+1/\alpha}\right)^{y}\left(\frac{1/\alpha}{\mu+1/\alpha}\right)^{1/\alpha} \\ &= \frac{\Gamma\left(y+1/\alpha\right)}{\Gamma\left(y+1\right)\Gamma\left(1/\alpha\right)}\left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y}\left(\frac{1}{1+\alpha\mu}\right)^{1/\alpha}, \quad y=0,1,\ldots, \end{aligned} \tag{1}$$

where the expected value $\mathrm{E}_{\mathsf{NB}}(Y;\mu)=\mu>0$ is the mean, $\alpha>0$ is the dispersion parameter, and where $\Gamma(\cdot)$ denotes the gamma function [14]. If Y has the PMF given in (1), then we simply write $Y\sim\mathcal{NB}(\mu,\alpha)$. When $\alpha\to 0^{+}$, the NB-D approaches a Pois-D with mean μ [11]. The variance and skewness of $Y\sim\mathcal{NB}(\mu,\alpha)$ are

$$\mathsf{Var}_{\mathsf{NB}}(Y;\mu,\alpha)=\mu+\alpha\mu^{2}, \tag{2}$$

and

$$\mathsf{Skew}_{\mathsf{NB}}(Y;\mu,\alpha)=\sqrt{\mu(1+\alpha\mu)}(1+2\alpha\mu),$$

while the kurtosis is given by

$$\mathsf{Kurt}_{\mathsf{NB}}(Y;\mu,\alpha)=3+\mathsf{ExKurt}_{\mathsf{N}}(Y;\mu,\alpha),$$

where

$$\mathsf{ExKurt}_{\mathsf{N}}(Y;\mu,\alpha)=6\alpha+\left(\mu+\alpha\mu^{2}\right)^{-1} \tag{3}$$

represents the excess kurtosis in comparison to the normal distribution.

As discussed in [14], even though the NB-D alleviates the highly restrictive assumption of equidispersion posed by the Pois-D, there are instances where NB-Ds may also be overdispersed. While Poisson overdispersion occurs when its observed distributional variance exceeds the mean, NB overdispersion occurs when the calculated model variance exceeds the nominal NB variance. The NB-D may, therefore, be inadequate in modelling the variance inherent in the data. It is thus possible that both Poisson and NB overdispersions might occur at the same time.

While one common cause of overdispersion is excess zeros by an additional data-generating process, another contributing factor to larger variances, and thus possible overdispersion, is the presence of outliers in the data [14]. As discussed by Ritter [29], real-world data is often contaminated with outliers or atypical observations that can affect the estimation of model parameters, or in the context of regression, the regression coefficients. Inappropriate imposition of the Pois-D and NB-D may underestimate the standard errors and overstate the significance of the regression coefficients, which could lead to misleading inference.

This raises the question: how should outliers be handled? To answer this, it is important to note that outliers are generally divided into two broad categories:

1. Mild outliers: observations sampled from some population different or even far from the assumed model.
2. Gross outliers: observations that cannot be modelled by a distribution as they are unpredictable.

In the presence of gross outliers, the recommended approach is to eliminate the observations or choose a suitable method for suppressing them [2]. For mild outliers, however, it is usually recommended to use a model flexible enough to accommodate the atypical data points [29, 26], which is of specific interest in this paper. We propose the contaminated NB distribution (cNB-D) on which the majority of observations are from the NB-D, and the minority proportion is from an NB-D with higher dispersion. This is analogous to works presented in [20, 21, 27, 28, 33, 35]. The resulting model is in the form of a two-component mixture, with one component representing the "good" observations (reference NB-D) and the other having the same mean but an inflated dispersion parameter, representing the "bad" observations (contaminant distribution), making it flexible enough to accommodate mild outliers. Note that both of these components have the same mean, which is the mean accounted for by the majority of the data that are considered "good"; additionally, this provides for a more parsimonious model. For a discussion on the concept of a reference distribution (which in this paper is assumed to be the NB in (1)) see [7, 13].

Another issue arises when modelling the kurtosis of data. Although the excess kurtosis of $Y \sim \mathcal{NB}(\mu, \alpha)$ given in (3) is allowed to vary between 0^+ (when $\alpha \to 0^+$ and $\mu \to \infty$) and infinity (when $\mu \to 0^+$), this does not mean that we have control over it. To clarify, suppose we use the method of moments to estimate μ and α. This involves comparing the sample mean and variance with the model's mean and variance given (2), and then solving for μ and α. However, in doing so, we cannot manipulate the kurtosis, nor can we ensure that the model's kurtosis matches the empirical (sample) kurtosis. Therefore, despite the range of the excess kurtosis extending to infinity, it is fixed for a pair (μ, α).

As a motivating example, Figure 1 displays data from a medical study conducted in Germany, illustrating the possible outliers in the number of doctor visits. Government spending on health care surged in Germany in the 1980s and 1990s, and, in an effort to curtail this expenditure, a major healthcare reform took place in 1997. The reform raised the co-payments by up to 200% and introduced upper limits on the reimbursement of physicians by state insurance. Patients were surveyed for the one-year panel (1996) before and the one-year panel (1998) after reform to assess whether the number of physician visits by patients declined. The dataset from the German Socio-Economic Panel [12] can be downloaded from the Journal of Applied Econometrics Data Archive. For the 1-year panel of 1998 of working women, we focus on a subset of this data, as utilized by [14, 17, 39], to examine the number of doctor visits of patients who claimed to be of bad health. As illustrated in Figure 1, there seems to be an excess of points at the sides of the support (an excess of zeros on the left, and some too large values on the right) which may cause overdispersion. In this health study, mild outliers might include patients who excessively visited the doctor, much more than expected, given a model; therefore they can be considered as outliers in response to the assumed model. These outliers have the potential to introduce bias into the estimates of the regression coefficients, distort inference, and result in an overestimation of the overdispersion parameter, consequently leading to an overestimation of standard errors.

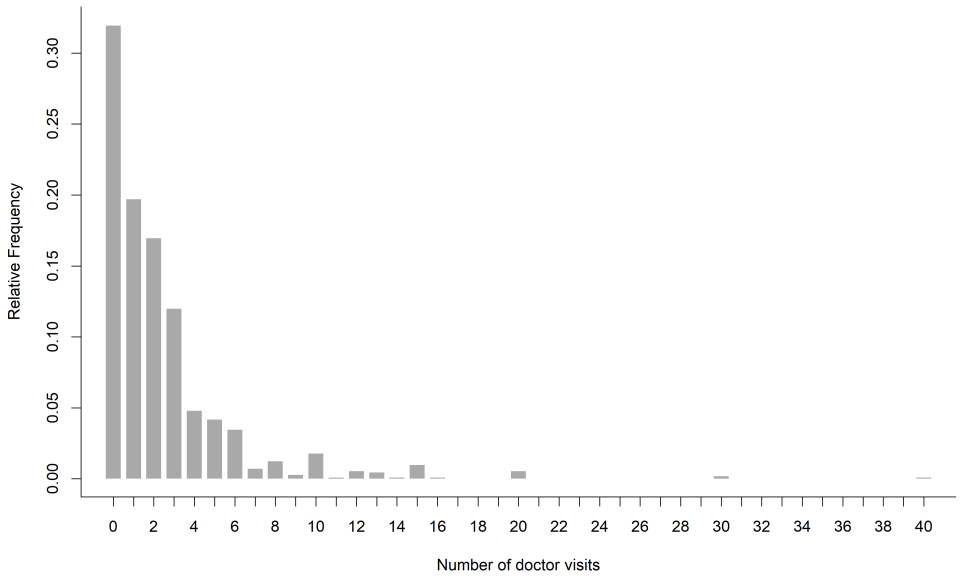


Figure 1: Barplot of the number of visits to a doctor in a year.

The paper is set out as follows. In Section 2, we construct the cNB regression model (cNB-RM) using the proposed cNB-D in a regression context, that can capture NB overdispersion and classify whether an observation is a mild outlier once the model is fitted. The corresponding expectation-maximization (EM) algorithm for maximum likelihood (ML) estimation is presented in Section 3.2, along with a discussion of proposed initialization strategies and the convergence criteria considered in Section 3.3. A simulation study in Section 4 illustrates its parameter recovery ability, followed by a sensitivity analysis to investigate the impact of a single atypical observation on the estimations. Two real-data applications are presented in Section 5 to illustrate the viability of the cNB-D as an alternative model for overdispersed data. Finally, conclusions are drawn in Section 6.

2. Methodological proposals

In this section, we introduce the cNB-D (Section 2.1) and the cNB-RM (Section 2.2).

2.1. The contaminated negative binomial model

The PMF of the proposed cNB-D is

$$f_{\mathsf{cNB}}(y;\mu,\alpha,\delta,\eta) = (1-\delta)\underbrace{f_{\mathsf{NB}}(y;\mu,\alpha)}_{\text{reference}} + \delta\underbrace{f_{\mathsf{NB}}(y;\mu,\eta\alpha)}_{\text{contaminant}}, \tag{4}$$

where $\delta \in (0,1)$ and $\eta > 1$. If Y has the PMF given in (4), then we will simply write $Y \sim c\mathcal{NB}(\mu,\alpha,\delta,\eta)$. As well-documented in [40], although not necessary, a restriction on δ can be imposed such that $\delta \in (0, 0.5)$. This will ensure that at least half of the sample points are considered "good", which is a general assumption within robust statistical inference. The additional contamination parameters δ and η have an interpretation of practical interest: δ is the proportion of points not from the reference distribution, while η denotes the degree of contamination. Since $\eta > 1$, it can be viewed as an inflation parameter, i.e., the increase in variability due to the points that do not come from the reference distribution. For example, if $\eta = 2$, then the dispersion of points from the contaminant NB component is twice that of the reference NB-D (as measured by α).

The moments of practical interest of $Y \sim c\mathcal{NB}(\mu,\alpha,\delta,\eta)$ are:

$$\mathrm{E}_{\mathsf{cNB}}(Y;\mu) = \mu,$$

$$\mathsf{Var}_{\mathsf{cNB}}(Y;\mu,\alpha,\delta,\eta) = \mu + [(1-\delta)+\delta\eta]\,\alpha\mu^2, \tag{5}$$

$$\begin{aligned}\mathsf{Skew}_{\mathsf{cNB}}(Y;\mu,\alpha,\delta,\eta) =& \frac{2\alpha^2\mu^3\left[(1-\delta)+\delta\eta^2\right] + 3\alpha\mu^2\left[(1-\delta)+\delta\eta\right] + \mu}{\sqrt{\alpha\mu^2\left[(1-\delta)+\delta\eta\right]+\mu}} \\ =& \mathsf{Skew}_{\mathsf{NB}}(Y;\mu,\alpha) + \left[\alpha\mu^2((1-\delta)+\delta\eta)+\mu\right]^{-\frac{1}{2}}\left[\mu\left(\alpha\mu\left(\delta(\eta-1)(2\alpha(\eta+1)\mu+3)\right.\right.\right. \\ & \left.\left.\left. -2\sqrt{(\alpha\mu+1)(\alpha\mu(\delta(\eta-1)+1)+1)}+2\alpha\mu+3\right) - \sqrt{(\alpha\mu+1)(\alpha\mu(\delta(\eta-1)+1)+1)}+1\right)\right].\end{aligned}$$

and

$$\mathsf{Kurt}_{\mathsf{cNB}}(Y;\mu,\alpha,\delta,\eta) = 3 + \mathsf{ExKurt}_{\mathsf{N}}(Y;\mu,\alpha) + \mathsf{ExKurt}_{\mathsf{NB}}(Y;\mu,\alpha,\delta,\eta),$$

where

$$\begin{aligned}\mathsf{ExKurt}_{\mathsf{NB}}(Y;\mu,\alpha,\delta,\eta) =& \left[(\alpha\mu+1)(\alpha\mu(\delta(\eta-1)+1)+1)^2\right]^{-1}\{\alpha\delta(\eta-1)\left(\alpha\mu\left(-\delta(\eta-1)\right.\right. \\ & \times(3(2\alpha+1)\mu(\alpha\mu+1)+1) + 6\alpha\eta^2\mu(\alpha\mu+1) + 3\eta(\alpha\mu+1)(2\alpha\mu+\mu+4) \\ & \left.\left. -3(2\alpha+1)\mu(\alpha\mu+1)+5\right)+5)\}\end{aligned} \tag{6}$$

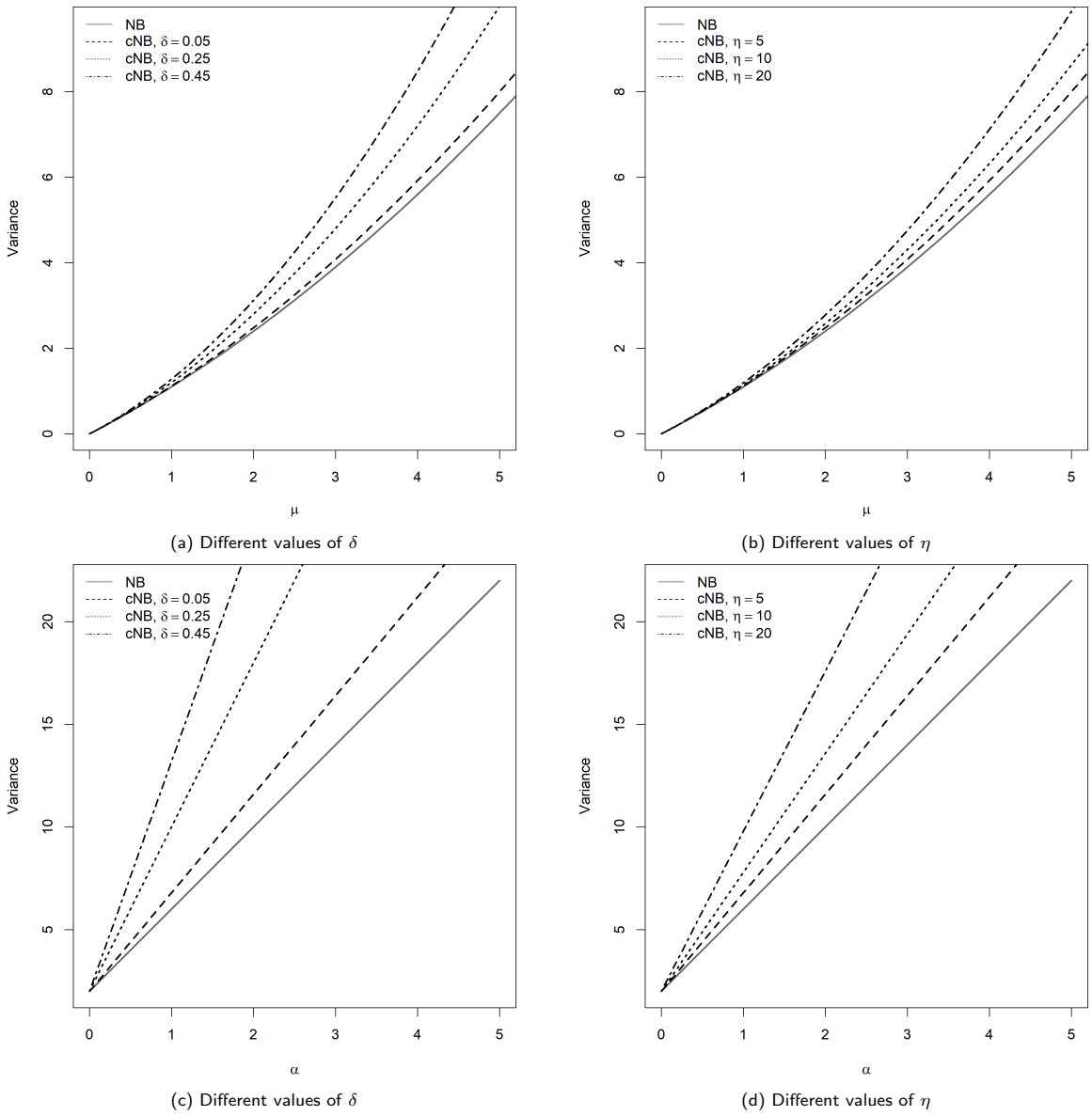
is the excess kurtosis in comparison to $\mathcal{NB}(\mu,\alpha)$. Since $[(1-\delta)+\delta\eta] > 1$, the variance in (5) exceeds that of the $\mathcal{NB}(\mu,\alpha)$ distribution, with the extent of the difference determined by the values of δ and η. Therefore, the cNB-D can alleviate possible NB overdispersion. Similarly, it can be shown that the numerator and denominator of the excess NB kurtosis in (6) are greater than 0. Consequently, the kurtosis of the $c\mathcal{NB}(\mu,\alpha,\delta,\eta)$ is larger than that of the $\mathcal{NB}(\mu,\alpha)$. In a similar way, the argument holds for skewness of the cNB-D. While not the primary focus of this paper, this valuable advantage of the cNB-D thus also includes specific characterization and control over the empirical skewness. Examples illustrating this are shown in Figures 2, 3, and 4 for different choices of δ and η.

The following proposition explores the limiting cases of the cNB-D at the border of its parameter space.

Proposition 1. *Let $Y \sim c\mathcal{NB}(\mu,\alpha,\delta,\eta)$, then:*

(a) if $\delta \to 0^+$, then $Y \xrightarrow{D} \mathcal{NB}(\mu,\alpha)$,

(b) if $\eta \to 1^+$, then $Y \xrightarrow{D} \mathcal{NB}(\mu,\alpha)$,



(a) Different values of δ

(b) Different values of η

(c) Different values of δ

(d) Different values of η

Figure 2: Examples illustrating the higher variance of the cNB-D (4) compared to the NB-D (1) for increasing values of μ (when $\alpha = 0.1$) and α (when $\mu = 2$), and for different values of δ (when $\eta = 5$) and η (when $\delta = 0.05$).

(c) *if* $\delta \to 0^+$ *and* $\alpha \to 0^+$, *then* $Y \xrightarrow{D} \mathcal{P}ois(\mu)$,

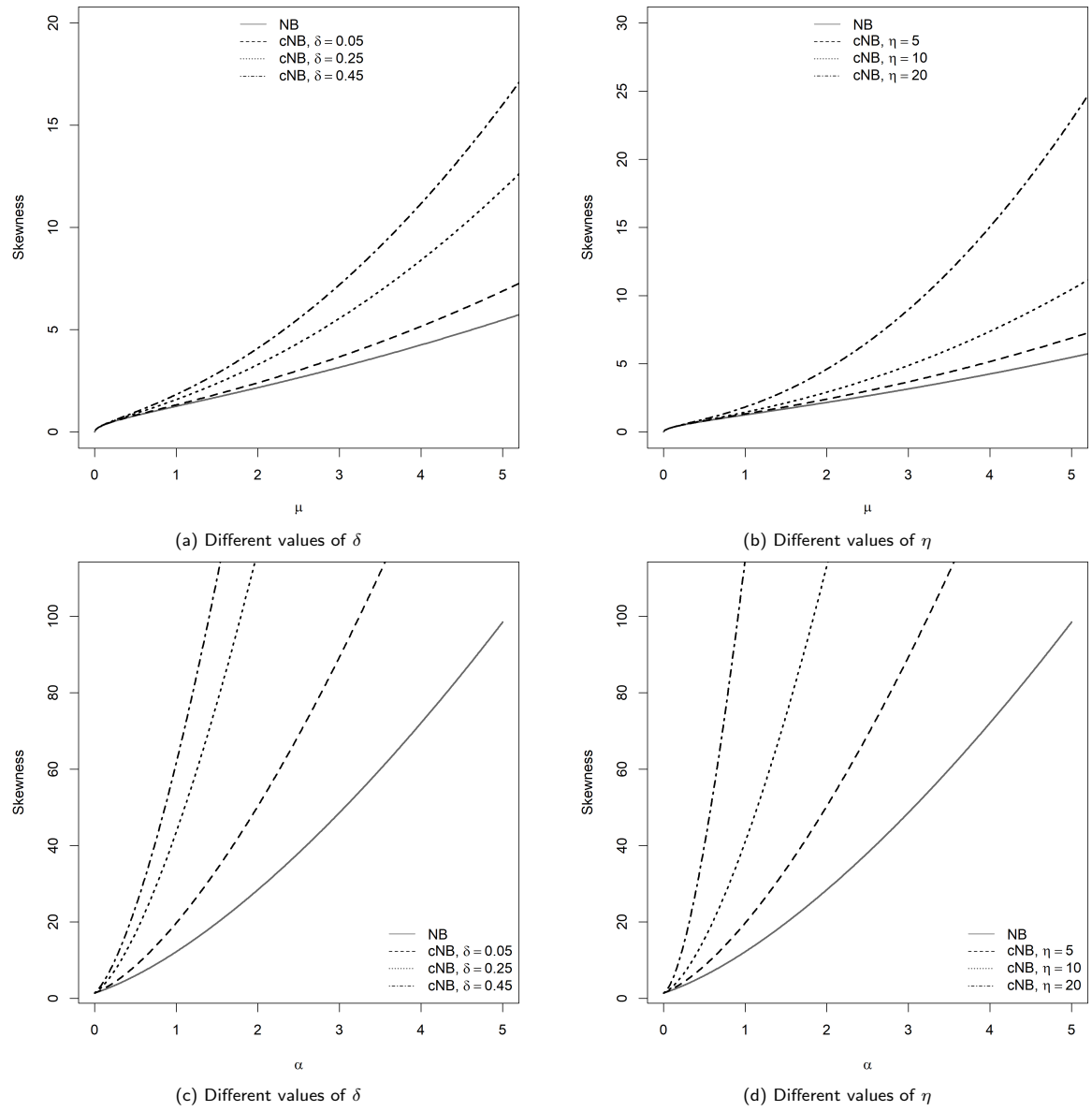
(d) *if* $\eta \to 1^+$ *and* $\alpha \to 0^+$, *then* $Y \xrightarrow{D} \mathcal{P}ois(\mu)$,

where $\xrightarrow{D}$ denotes convergence in distribution and $\mathcal{P}\text{ois}(\mu)$ denotes a Poisson distribution with mean μ.

Proof. See Appendix A. □

2.2. The contaminated negative binomial regression model

A regression model based on the cNB-D (4) follows by conditioning the distribution of the count response $Y_i, i = 1, \ldots, n$, on a $(k+1)$-dimensional vector of covariates $\boldsymbol{x}_i' = (1, x_{1i}, \ldots, x_{ki})$, and by considering a vector of regression coefficients $\boldsymbol{\beta} : (k+1) \times 1$ such that the expected count for $Y_i | \boldsymbol{X} = \boldsymbol{x}_i$, say $\mu(\boldsymbol{x}_i; \boldsymbol{\beta})$, is related



(a) Different values of δ

(b) Different values of η

(c) Different values of δ

(d) Different values of η

Figure 3: Examples illustrating higher skewness of the cNB-D (4) compared to the NB-D (1) for increasing values of μ (when $\alpha = 0.1$) and α (when $\mu = 2$), and for different values of δ (when $\eta = 5$) and η (when $\delta = 0.05$).

to the covariates $\boldsymbol{x}_i$ through a linear predictor, with paramaters $\boldsymbol{\beta}$, using a convenient link function. The log function is commonly used as the link function for count data. It thus follows

$$g(\mu(\boldsymbol{x}_i;\boldsymbol{\beta})) = \log\left(\mu(\boldsymbol{x}_i;\boldsymbol{\beta})\right) = \boldsymbol{x}_i'\boldsymbol{\beta},$$

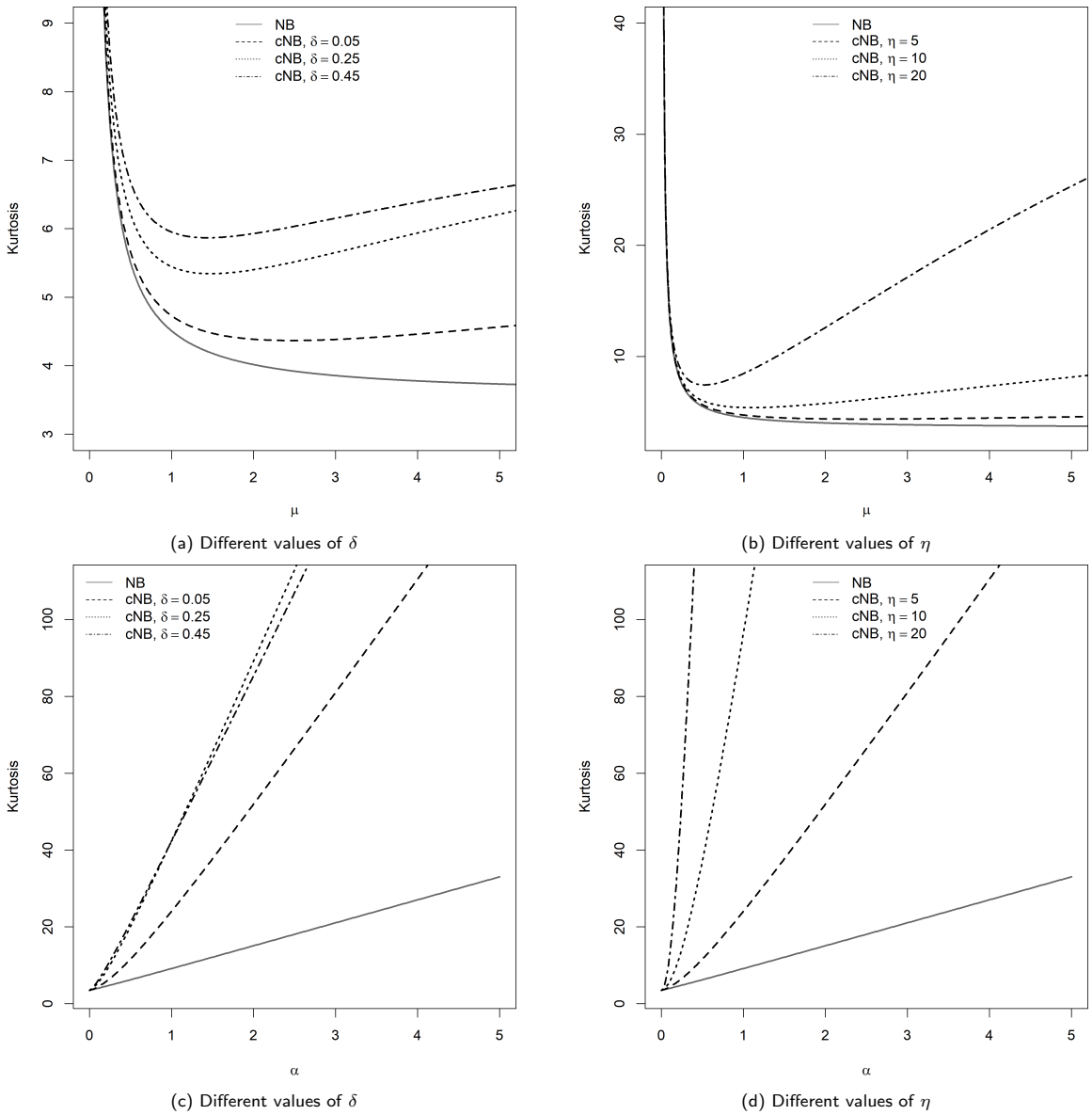
where g is the log link function. The inverse g^{-1} of the link function leads to

$$\mu(\boldsymbol{x}_i;\boldsymbol{\beta}) = E\left(Y_i|\boldsymbol{x}_i;\boldsymbol{\beta}\right) = g^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta}) = \mathrm{e}^{\boldsymbol{x}_i'\boldsymbol{\beta}}.$$

Thus, the PMF of $Y_i|\boldsymbol{x}_i$, according to the cNB-RM, is

$$f_{\mathsf{cNB}}(y_i;\mu(\boldsymbol{x}_i;\boldsymbol{\beta}),\alpha,\delta,\eta) = (1-\delta)f_{\mathsf{NB}}(y_i;\mu(\boldsymbol{x}_i;\boldsymbol{\beta}),\alpha) + \delta f_{\mathsf{NB}}(y_i;\mu(\boldsymbol{x}_i;\boldsymbol{\beta}),\eta\alpha). \tag{7}$$

If $\beta_1 = \cdots = \beta_k = 0$, then we have $Y_i|\boldsymbol{x}_i \sim c\mathcal{NB}(\mu = \mathrm{e}^{\beta_0},\alpha,\delta,\eta)$. It might be worth reiterating that, for each $\boldsymbol{x}_i$, (7) is in the form of a two-component mixture, where both components have the same mean $\mu(\boldsymbol{x}_i;\boldsymbol{\beta})$, but



(a) Different values of δ

(b) Different values of η

(c) Different values of δ

(d) Different values of η

Figure 4: Examples illustrating higher kurtosis of the cNB-D (4) compared to the NB-D (1) for increasing values of μ (when $\alpha = 0.1$) and α (when $\mu = 2$), and for different values of δ (when $\eta = 5$) and η (when $\delta = 0.05$).

one component has an inflated variance. An advantage of model (7) is that, given the estimates of $\boldsymbol{\beta}, \alpha, \delta$, and η, it is possible to determine whether a data point $(\boldsymbol{x}_i, y_i)$ is good or not, with respect to the reference NB-D, via the *a posteriori* probability

$$P\left((\boldsymbol{x}_i, y_i) \text{ comes from the reference NB-RM}|\hat{\boldsymbol{\beta}}, \hat{\alpha}, \hat{\delta}, \hat{\eta}\right) = \frac{(1-\hat{\delta}) f_{\text{NB}}(y_i; \mu(\boldsymbol{x}_i; \hat{\boldsymbol{\beta}}), \hat{\alpha})}{f_{\text{cNB}}(y_i; \mu(\boldsymbol{x}_i; \hat{\boldsymbol{\beta}}), \hat{\alpha}, \hat{\delta}, \hat{\eta})}. \tag{8}$$

It is natural to consider $(\boldsymbol{x}_i, y_i)$ as "good" if the probability in (8) is greater than 0.5, and a (mild) outlier otherwise.

3. Inference by maximum likelihood: application of the EM algorithm

In this section, we discuss the identifiability of the cNB-D (Section 3.1), followed by a presentation of the EM algorithm (Section 3.2) for ML estimation of the more general cNB-RM. The initialization strategy and

convergence criteria considered (Section 3.3) and an explanation of how the standard errors of the estimates are computed (Section 3.4) are also discussed.

3.1. Identifiability

Identifiability is a fundamental prerequisite for many statistical procedures, including the asymptotic theory in ML estimation of model parameters. In [37], it is demonstrated that finite NB-mixtures are identifiable. This result is particularly relevant since the cNB-D can be viewed as a two-component NB-mixture, with both components sharing the same mean parameter μ. Consequently, the cNB-D inherits the identifiability properties of the NB mixture model. This guarantees that the model parameters can be uniquely determined, thereby underpinning the reliability of subsequent statistical inference and predictions based on the model.

3.2. An EM algorithm

We illustrate the EM algorithm for the more general cNB-RM. Let $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ be an observed sample from the cNB-RM (7). For the application of the EM algorithm, it is convenient to view the observed data as incomplete. In this case, the source of incompleteness stems from the fact that we do not know if the generic data point $(\boldsymbol{x}_i, y_i)$ is an outlier. To denote the source of incompleteness, we use an indicator vector $\boldsymbol{v} = (v_1, \ldots, v_n)$ so that $v_i = 1$ if $(\boldsymbol{x}_i, y_i)$ is a mild outlier (does not come from the reference distribution) and $v_i = 0$ otherwise. The complete-data are thus given by $(\boldsymbol{x}_1, y_1, v_1), \ldots, (\boldsymbol{x}_n, y_n, v_n)$ and, from (1) and (7), the complete-data likelihood function can be written as

$$\begin{aligned}
L_c(\boldsymbol{\beta}, \alpha, \delta, \eta) &= \prod_{i=1}^{n} \left[(1-\delta) f_{\mathsf{NB}}(y_i, \mu(\boldsymbol{x}_i;\boldsymbol{\beta}), \alpha)\right]^{1-v_i} \left[\delta f_{\mathsf{NB}}(y_i; \mu(\boldsymbol{x}_i;\boldsymbol{\beta}), \eta\alpha)\right]^{v_i} \\
&= \prod_{i=1}^{n} \left[\frac{(1-\delta)\Gamma\left(y_i + 1/\alpha\right)}{\Gamma\left(y_i+1\right)\Gamma\left(1/\alpha\right)} \left(\frac{\alpha\mu(\boldsymbol{x}_i;\boldsymbol{\beta})}{1+\alpha\mu(\boldsymbol{x}_i;\boldsymbol{\beta})}\right)^{y_i} \left(\frac{1}{1+\alpha\mu(\boldsymbol{x}_i;\boldsymbol{\beta})}\right)^{1/\alpha} \right]^{1-v_i} \\
&\quad \times \left[\frac{\delta\Gamma\left(y_i + 1/(\eta\alpha)\right)}{\Gamma\left(y_i+1\right)\Gamma\left(1/(\eta\alpha)\right)} \left(\frac{(\eta\alpha)\mu(\boldsymbol{x}_i;\boldsymbol{\beta})}{1+(\eta\alpha)\mu(\boldsymbol{x}_i;\boldsymbol{\beta})}\right)^{y_i} \left(\frac{1}{1+(\eta\alpha)\mu(\boldsymbol{x}_i;\boldsymbol{\beta})}\right)^{1/(\eta\alpha)} \right]^{v_i}.
\end{aligned}$$

The complete-data log-likelihood function then follows as

$$l_c(\boldsymbol{\beta}, \alpha, \delta, \eta) = l_{c_1}(\delta) + l_{c_2}(\boldsymbol{\beta}, \alpha, \eta), \tag{9}$$

where

$$l_{c_1}(\delta) = \sum_{i=1}^{n} (1-v_i)\ln(1-\delta) + v_i \ln\delta$$

and

$$\begin{aligned}
&l_{c_2}(\boldsymbol{\beta}, \alpha, \eta) \\
&= \sum_{i=1}^{n} (1-v_i) \left[\ln\Gamma\left(y_i + \frac{1}{\alpha}\right) - \ln\Gamma(y_i+1) - \ln\Gamma\left(\frac{1}{\alpha}\right) + y_i \ln(\alpha\mu(\boldsymbol{x}_i;\boldsymbol{\beta})) - \left(y_i + \frac{1}{\alpha}\right)\ln(1+\alpha\mu(\boldsymbol{x}_i;\boldsymbol{\beta})) \right] \\
&\quad + v_i \left[\ln\Gamma\left(y_i + \frac{1}{\eta\alpha}\right) - \ln\Gamma(y_i+1) - \ln\Gamma\left(\frac{1}{\eta\alpha}\right) + y_i \ln(\eta\alpha\mu(\boldsymbol{x}_i;\boldsymbol{\beta})) - \left(y_i + \frac{1}{\eta\alpha}\right)\ln(1+\eta\alpha\mu(\boldsymbol{x}_i;\boldsymbol{\beta})) \right].
\end{aligned}$$

Alternatively, the log-likelihood in (9) can be written in terms of the model coefficients as

$$\begin{aligned}
&l_c(\boldsymbol{\beta}, \alpha, \delta, \eta) \\
&= \sum_{i=1}^{n} (1-v_i) \left[\ln(1-\delta) + \ln\Gamma\left(y_i + \frac{1}{\alpha}\right) - \ln\Gamma(y_i+1) - \ln\Gamma\left(\frac{1}{\alpha}\right) + y_i \ln\left(\alpha \mathrm{e}^{\boldsymbol{x}_i'\boldsymbol{\beta}}\right) - \left(y_i + \frac{1}{\alpha}\right)\ln\left(1+\alpha \mathrm{e}^{\boldsymbol{x}_i'\boldsymbol{\beta}}\right) \right] \\
&\quad + v_i \left[\ln\delta + \ln\Gamma\left(y_i + \frac{1}{\eta\alpha}\right) - \ln\Gamma(y_i+1) - \ln\Gamma\left(\frac{1}{\eta\alpha}\right) + y_i \ln\left(\eta\alpha \mathrm{e}^{\boldsymbol{x}_i'\boldsymbol{\beta}}\right) - \left(y_i + \frac{1}{\eta\alpha}\right)\ln\left(1+\eta\alpha \mathrm{e}^{\boldsymbol{x}_i'\boldsymbol{\beta}}\right) \right].
\end{aligned}$$

The algorithm iterates between the E-step and M-step until convergence. These steps for the $(r+1)$th iteration of the algorithm are detailed below.

E-step

In the E-step, we compute the conditional expectation of the complete-data log-likelihood function as

$$Q\left(\boldsymbol{\beta},\alpha,\delta,\eta|\boldsymbol{\beta}^{(r)},\alpha^{(r)},\delta^{(r)},\eta^{(r)}\right)=Q_1\left(\delta|\boldsymbol{\beta}^{(r)},\alpha^{(r)},\delta^{(r)},\eta^{(r)}\right)+Q_2\left(\boldsymbol{\beta},\alpha,\eta|\boldsymbol{\beta}^{(r)},\alpha^{(r)},\delta^{(r)},\eta^{(r)}\right)$$

for the $(r+1)$-th iteration, which is in the same order as (9). $Q\left(\boldsymbol{\beta},\alpha,\delta,\eta|\boldsymbol{\beta}^{(r)},\alpha^{(r)},\delta^{(r)},\eta^{(r)}\right)$ is obtained by substituting v_i in (9) by the expected a posteriori probability for a point to be an outlier

$$\mathrm{E}\left(V_i|y_i,\boldsymbol{x}_i;\boldsymbol{\beta}^{(r)},\alpha^{(r)},\delta^{(r)},\eta^{(r)}\right)=\frac{\delta^{(r)}f_{\mathsf{NB}}\left(y_i|\boldsymbol{x}_i;\mu\left(\boldsymbol{x}_i;\boldsymbol{\beta}^{(r)}\right),\eta^{(r)}\alpha^{(r)}\right)}{f_{\mathsf{cNB}}\left(y_i|\boldsymbol{x}_i;\mu\left(\boldsymbol{x}_i;\boldsymbol{\beta}^{(r)}\right),\alpha^{(r)},\delta^{(r)},\eta^{(r)}\right)}:=v_i^{(r)}. \tag{10}$$

M-step

An update $\delta^{(r+1)}$ for δ is calculated by independently maximizing

$$Q_1\left(\delta|\boldsymbol{\beta}^{(r)},\alpha^{(r)},\delta^{(r)},\eta^{(r)}\right)=\sum_{i=1}^{n}\left\{\left(1-v_i^{(r)}\right)\ln\left(1-\delta\right)+v_i^{(r)}\ln\delta\right\}$$

with respect to δ and subject to the constraints on this parameter. It follows that

$$\delta^{(r+1)}=\frac{1}{n}\sum_{i=1}^{n}v_i^{(r)}.$$

If we assume $\delta<0.5$, then

$$\delta^{(r+1)}=\min\left\{0.5,\frac{1}{n}\sum_{i=1}^{n}v_i^{(r)}\right\}.$$

Updates for $\boldsymbol{\beta}$, α, and η are obtained by maximizing

$$\begin{aligned}
&Q_2\left(\boldsymbol{\beta},\alpha,\eta|\boldsymbol{\beta}^{(r)},\alpha^{(r)},\delta^{(r)},\eta^{(r)}\right)\\
=&\sum_{i=1}^{n}\left(1-v_i^{(r)}\right)\left[\ln\Gamma\left(y_i+\frac{1}{\alpha}\right)-\ln\Gamma(y_i+1)-\ln\Gamma\left(\frac{1}{\alpha}\right)+y_i\ln\left(\alpha\mu\left(\boldsymbol{x}_i;\boldsymbol{\beta}\right)\right)\right.\\
&\left.-\left(y_i+\frac{1}{\alpha}\right)\ln\left(1+\alpha\mu\left(\boldsymbol{x}_i;\boldsymbol{\beta}\right)\right)\right]+v_i\left[\ln\Gamma\left(y_i+\frac{1}{\eta\alpha}\right)-\ln\Gamma(y_i+1)-\ln\Gamma\left(\frac{1}{\eta\alpha}\right)\right.\\
&\left.+y_i\ln\left(\eta\alpha\mu(\boldsymbol{x}_i;\boldsymbol{\beta})\right)-\left(y_i+\frac{1}{\eta\alpha}\right)\ln\left(1+\eta\alpha\mu\left(\boldsymbol{x}_i;\boldsymbol{\beta}\right)\right)\right]
\end{aligned}$$

This can be achieved in `R` using the `optim()` function included in the **stats** package. The BFGS algorithm, which is used for unconstrained optimization, can be passed to `optim()` via the `method` argument. Since some of the parameters involved have constraints, the following transformations/back-transformations are implemented:

$$\begin{aligned}
\tilde{\alpha}=\ln(\alpha)&\longleftrightarrow\alpha=\mathrm{e}^{\tilde{\alpha}}\\
\tilde{\eta}=\ln(\eta-1)&\longleftrightarrow\eta=\mathrm{e}^{\tilde{\eta}}+1,
\end{aligned}$$

where parameters marked with a 'tilde' denote the unconstrained parameters.

3.3. Initialization and convergence

The starting values are a critical step in EM-based algorithms and can greatly impact the accuracy and reliability of the model estimates [5, 27]; their choice thus constitutes an important aspect of estimation. If the starting values are chosen poorly, the algorithm may converge to a local maximum instead of the global maximum. Moreover, if the starting values deviate too much from the true values, the algorithm may converge slowly or not at all. We suggest fitting a standard NB-RM using the same predictors that would be used to fit a cNB-RM to

the data. The estimated coefficients can then serve as initial values for fitting a cNB-RM. For δ and η, we suggest choosing them such that the cNB-RM tends to the NB-RM, i.e., $\delta^{(0)} \to 1^-$ (or $\delta^{(0)} \to 0^+$) and $\eta^{(0)} \to 1^+$.

As for the stopping rule, there are several convergence criteria that can be used to determine whether the EM algorithm has converged or not. One common method is to track the change in the observed-data log-likelihood function, say l, between consecutive iterations. If the change falls below a predetermined threshold ϵ the algorithm can be considered to be converged, i.e., $l\left(\mu^{(r+1)}, \alpha^{(r+1)}, \delta^{(r+1)}, \eta^{(r+1)}\right) - l\left(\mu^{(r)}, \alpha^{(r)}, \delta^{(r)}, \eta^{(r)}\right) < \epsilon$. Due to the possibility of flat likelihoods, we employ stopping criteria of $\epsilon = 1 \times 10^{-10}$ or 1000 iterations.

3.4. Standard errors of the estimates

After executing the EM algorithm described in Section 3.2, the variance-covariance matrix of the parameter estimates is obtained by inverting the negative Hessian matrix, computed using the `optim()` function in `R`. The standard errors of the parameter estimates of the cNB-RM are then calculated as the square roots of the diagonal elements of this variance-covariance matrix.

4. Simulation studies

In this section, various aspects related to the proposed cNB-RM are investigated. We only focus on the cNB-RM because it is more general than the cNB-D. As the EM algorithm is used to fit the model, assessing its performance in terms of parameter recovery is imperative. The results of a parameter recovery study are presented in Section 4.1, while a sensitivity analysis that aims to evaluate the influence of a single outlier on the parameter estimates of the NB-RM and cNB-RM is illustrated in Section 4.2. The ability of the cNB-RM to determine whether an observation is "good" or "bad", as given in (8), is also evaluated.

4.1. Parameter recovery

Parameter recovery focuses on the algorithm's ability to accurately retrieve the true generating parameters. If, across multiple replications, the mean of the estimates significantly deviates from the actual generating parameter, the estimator is deemed biased. Additionally, the extent of variability in the estimates across these replications is a matter of concern. With small to moderate sample sizes, the ML estimator of the dispersion parameter may be subject to significant bias, which, in turn, may affect the coefficient estimates [16]. The bias may even be more pronounced in the case of the cNB-D. If only a few mild outliers are in the sample, this could lead to sparse data for the contaminated component in (4) leading to numerical instabilities in η and δ.

As described in [14], the NB-D is typically derived as a Poisson-gamma mixture. Although this result can be exploited in the generation of NB data, we directly used the `rnbinom()` function in the **stats** package. Since the cNB-D is a specific instance of a two-component NB mixture, the Bernoulli random variable V with probability of "success" δ (as introduced in Section 3.2 for the EM algorithm) can be used to select from which component (the reference or the contaminated NB components) each observation is generated.

We simulate 1000 samples with sizes $n = 100, 500,$ and 1000 to assess the accuracy of the point estimates of the EM algorithm described in Section 3.2. To generate data, we consider the following scenarios:

1. cNB-RM with $\alpha = 0.5$ and $\eta = 2$ for $\delta = \{0.05, 0.45\}$,
2. cNB-RM with $\alpha = 0.5$ and $\eta = 8$ for $\delta = \{0.05, 0.45\}$,

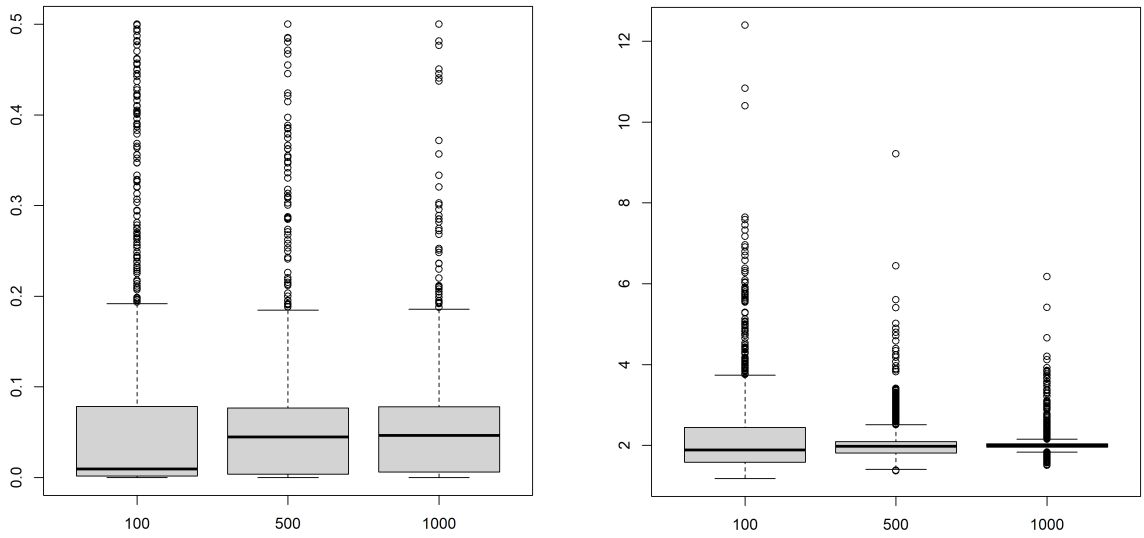
with an intercept of $\beta_0 = 20$, a binary covariate generated from a Bernoulli distribution with a probability of success of 0.5, and a continuous covariate generated by a uniform distribution over the interval $(-1, 1)$, with coefficients $\beta_1 = 0.75$ and $\beta_2 = -1.5$, respectively. In each scenario, we fit the cNB-RM to the generated data. For comparison's sake, the bias, and mean squared error (MSE) of the estimates, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$, $\hat{\alpha}$, $\hat{\delta}$, and $\hat{\eta}$, are reported in Tables 1 and 2. Additionally, boxplots of $\hat{\delta}$ and $\hat{\eta}$ are displayed for each scenario.

Table 1: Parameter recovery results for Scenario 1 with an intercept of $\beta_0 = 20$, a binary covariate with coefficient $\beta_1 = 0.75$, and a continuous covariate with coefficient $\beta_2 = -1.5$, $\alpha = 0.5$ and $\eta = 2$ for different values of δ, based on 1000 replications of varying sample sizes.

		$\delta = 0.05$			$\delta = 0.45$		
		$n = 100$	$n = 500$	$n = 1000$	$n = 100$	$n = 500$	$n = 1000$
Bias	$\hat{\beta}_0$	-0.0042	-0.0015	-0.0002	-0.0135	-0.0037	-0.0011
	$\hat{\beta}_1$	-0.0049	-0.0007	-0.0015	0.0089	-0.0009	-0.0005
	$\hat{\beta}_2$	0.0028	-0.0002	0.0013	-0.0063	0.0027	-0.0001
	$\hat{\alpha}$	-0.0434	-0.0147	-0.0076	-0.0053	0.0138	0.0096
	$\hat{\delta}$	0.0384	0.0188	0.0122	-0.0220	-0.0284	-0.0227
	$\hat{\eta}$	0.2815	0.0864	0.0736	0.4583	0.0958	0.0525
MSE	$\hat{\beta}_0$	0.0106	0.0022	0.0010	0.0151	0.0028	0.0014
	$\hat{\beta}_1$	0.0238	0.0047	0.0020	0.0300	0.0055	0.0028
	$\hat{\beta}_2$	0.0160	0.0031	0.0015	0.0243	0.0043	0.0022
	$\hat{\alpha}$	0.0143	0.0032	0.0017	0.0198	0.0061	0.0033
	$\hat{\delta}$	0.0235	0.0107	0.0071	0.0130	0.0142	0.0106
	$\hat{\eta}$	1.5117	0.3185	0.1368	2.4028	0.2808	0.1202

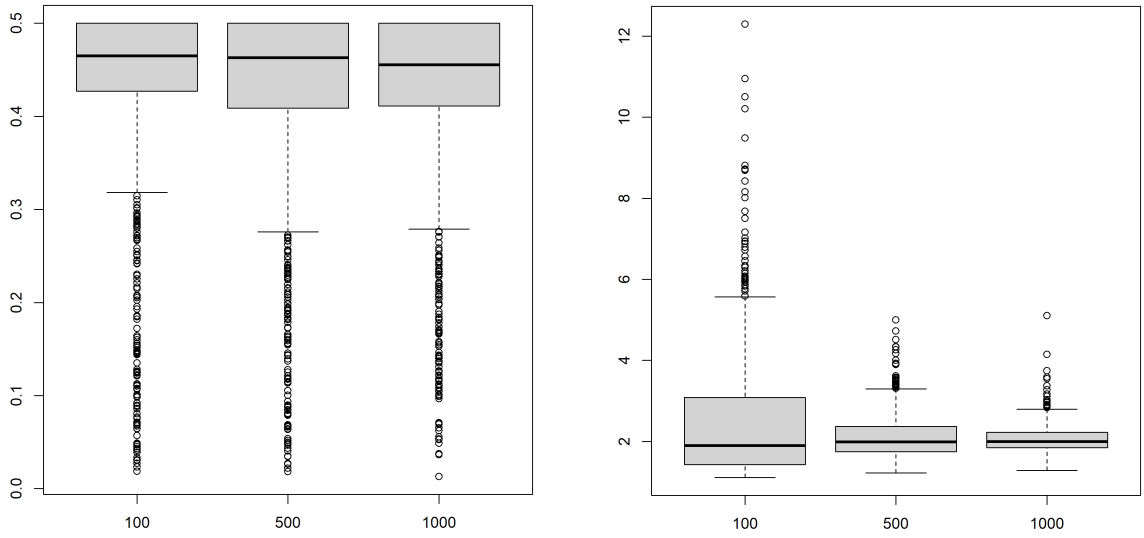
Table 2: Parameter recovery results for Scenario 2 with an intercept of $\beta_0 = 20$, a binary covariate with coefficient $\beta_1 = 0.75$, and a continuous covariate with coefficient $\beta_2 = -1.5$, $\alpha = 0.5$ and $\eta = 8$ for different values of δ, based on 1000 replications of varying sample sizes.

		$\delta = 0.05$			$\delta = 0.45$		
		$n = 100$	$n = 500$	$n = 1000$	$n = 100$	$n = 500$	$n = 1000$
Bias	$\hat{\beta}_0$	-0.0065	-0.0023	-0.0012	-0.0139	-0.0014	-0.0035
	$\hat{\beta}_1$	-0.0030	0.0023	0.0000	-0.0024	0.0009	0.0001
	$\hat{\beta}_2$	-0.0029	-0.0001	-0.0001	0.0053	0.0017	-0.0001
	$\hat{\alpha}$	-0.0431	-0.0093	-0.0047	0.0221	0.0051	-0.0030
	$\hat{\delta}$	0.0469	0.0087	0.0038	-0.0316	-0.0088	-0.0023
	$\hat{\eta}$	0.2787	0.2393	0.1192	1.4173	0.1979	0.1682
MSE	$\hat{\beta}_0$	0.0118	0.0023	0.0010	0.0282	0.0051	0.0026
	$\hat{\beta}_1$	0.0233	0.0046	0.0023	0.0548	0.0098	0.0050
	$\hat{\beta}_2$	0.0190	0.0036	0.0017	0.0409	0.0072	0.0037
	$\hat{\alpha}$	0.0126	0.0017	0.0008	0.0482	0.0072	0.0038
	$\hat{\delta}$	0.0140	0.0011	0.0004	0.0093	0.0024	0.0014
	$\hat{\eta}$	68.4501	10.3471	3.7658	20.5134	1.5495	0.8231



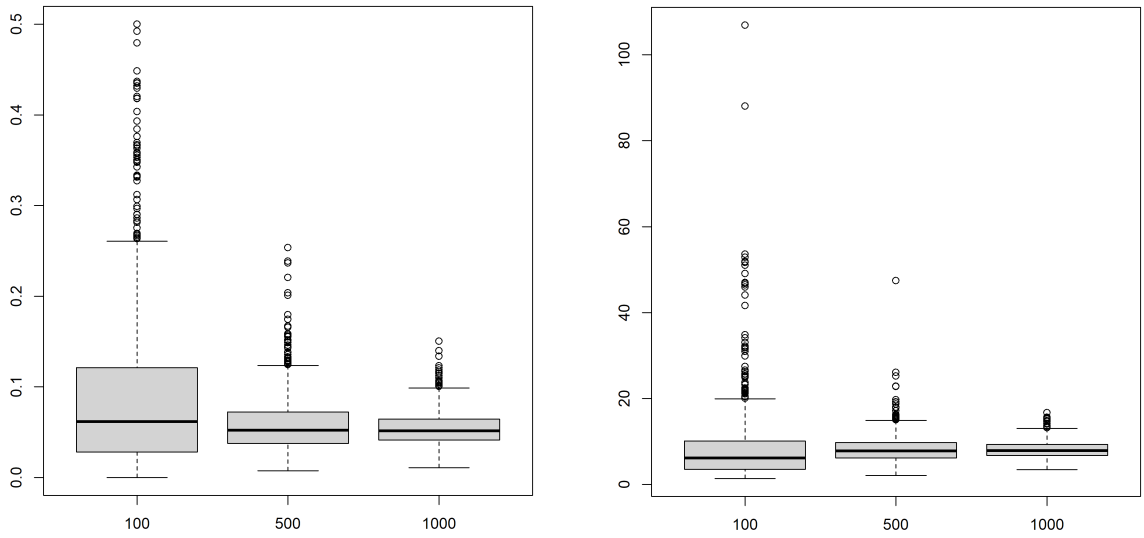
(a) $\hat{\delta}$ (b) $\hat{\eta}$

Figure 5: Boxplot of $\hat{\delta}$ and $\hat{\eta}$ for Scenario 1: cNB-RM with an intercept of $\beta_0 = 20$, a binary covariate with coefficient $\beta_1 = 0.75$, and a continuous covariate with coefficient $\beta_2 = -1.5$, $\alpha = 0.5$, $\delta = 0.05$ and $\eta = 2$, for 1000 replications of sample size $n = 100, 500, 1000$.



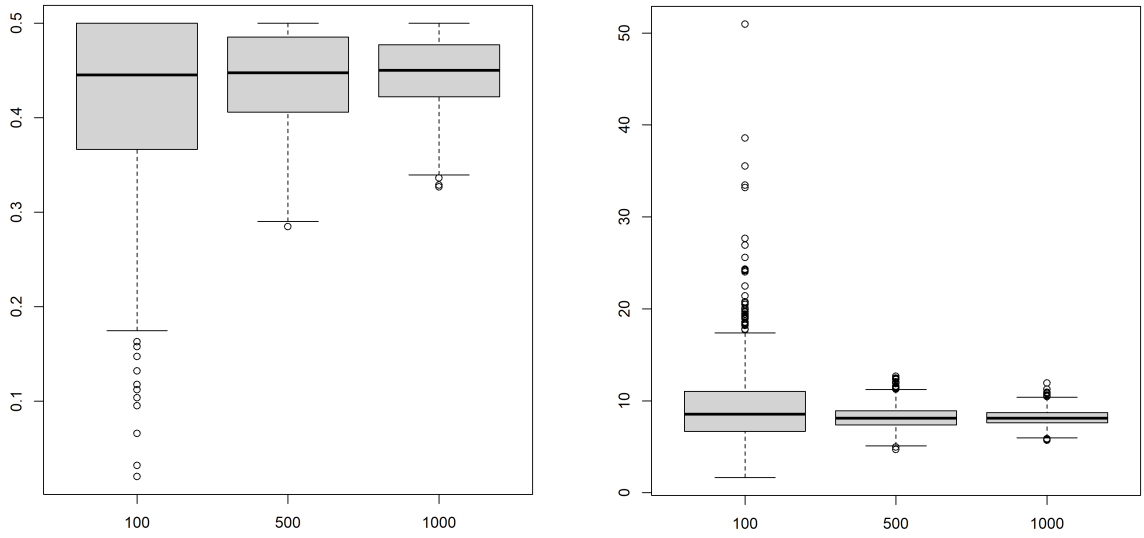
(a) $\hat{\delta}$ (b) $\hat{\eta}$

Figure 6: Boxplots of $\hat{\delta}$ and $\hat{\eta}$ for Scenario 1: cNB-RM with an intercept of $\beta_0 = 20$, a binary covariate with coefficient $\beta_1 = 0.75$, and a continuous covariate with coefficient $\beta_2 = -1.5$, $\alpha = 0.5$, $\delta = 0.45$ and $\eta = 2$, for 1000 replications of sample size $n = 100, 500, 1000$.



(a) $\hat{\delta}$ (b) $\hat{\eta}$

Figure 7: Boxplots of $\hat{\delta}$ and $\hat{\eta}$ for Scenario 2: cNB-RM with an intercept of $\beta_0 = 20$, a binary covariate with coefficient $\beta_1 = 0.75$, and a continuous covariate with coefficient $\beta_2 = -1.5$, $\alpha = 0.5$, $\delta = 0.05$ and $\eta = 8$, for 1000 replications of sample size $n = 100, 500, 1000$.



(a) $\hat{\delta}$ (b) $\hat{\eta}$

Figure 8: Boxplots of $\hat{\delta}$ and $\hat{\eta}$ for Scenario 2: cNB-RM with an intercept of $\beta_0 = 20$, a binary covariate with coefficient $\beta_1 = 0.75$, and a continuous covariate with coefficient $\beta_2 = -1.5$, $\alpha = 0.5$, $\delta = 0.45$ and $\eta = 8$, for 1000 replications of sample size $n = 100, 500, 1000$.

In the various scenarios investigated, we observe accurate parameter recovery and note that an increase in sample size leads to reduced bias, variability, and consequently MSE, within these estimates. The estimation of the regression coefficients is nearly perfect across all the data configurations. Instead, η and δ, which are the tailedness parameters of the model, are more difficult to estimate in the following sense. Tailedness parameters govern the tail behavior of the distribution, and, inferentially speaking, their estimation is primarily based on a small portion of the data, specifically those observations located in the tails. Because of this, comparing the performance of their estimates with those of other parameters, such as μ and α in the cNB-D, is not entirely fair, as it involves comparing the quality of estimates based on a different number of data points. Consequently, from an asymptotic perspective, the maximum likelihood (ML) estimators of the tailedness parameters typically require a larger sample size n to ensure the classical convergence properties of ML estimators, and this is evident in the results of both scenarios. For more details on this issue, see, for example, [25, 34, 32, 36]. Specifically, the MSEs consistently exhibit larger values when $\delta = 0.05$ compared to the case when $\delta = 0.45$. In light of this reasoning, this then makes intuitive sense: when $n = 100$, for example, only about 5 observations contribute to estimating η under $\delta = 0.05$, whereas approximately 45 observations are utilized when $\delta = 0.45$. This is even more pronounced when the proportion of outliers is small and the degree of contamination is high, as seen when comparing $\eta = 2$ and $\eta = 8$ for $\delta = 0.05$.

4.2. Sensitivity analysis

In this study, we perform a sensitivity analysis to investigate the impact of a single atypical observation on the Poisson-RM, NB-RM, and cNB-RM. We generate datasets of size $n = 200$ from the NB-RM with an intercept of $\beta_0 = 2$, and a continuous covariate generated by a uniform distribution over the interval $(-1, 1)$ with coefficient $\beta_1 = 1$, and $\alpha = 0.1$. A single outlier is then added to the generated data using one of the following schemes:

1. Close: A response value $y = 20$ close enough to the generated values and the predictor $x = -0.5$.
2. Far: A response value $y = 30$ far from the generated values and the predictor $x = -0.5$.

An example of a dataset generated with the above schemes is given in Figure 9, where the outlier is added to the generated NB data and illustrated in green for the close case and red for the far case. For each configuration, $n = 1000$ response counts are generated from an NB-D along with the covariate x being generated from a uniform distribution over the interval $(-1, 1)$. The Poisson-RM, NB-RM, and cNB-RM are then fit to the data to see what the impact of the outlier is on the estimated regression coefficients. The bias and MSE are reported in Table 3.

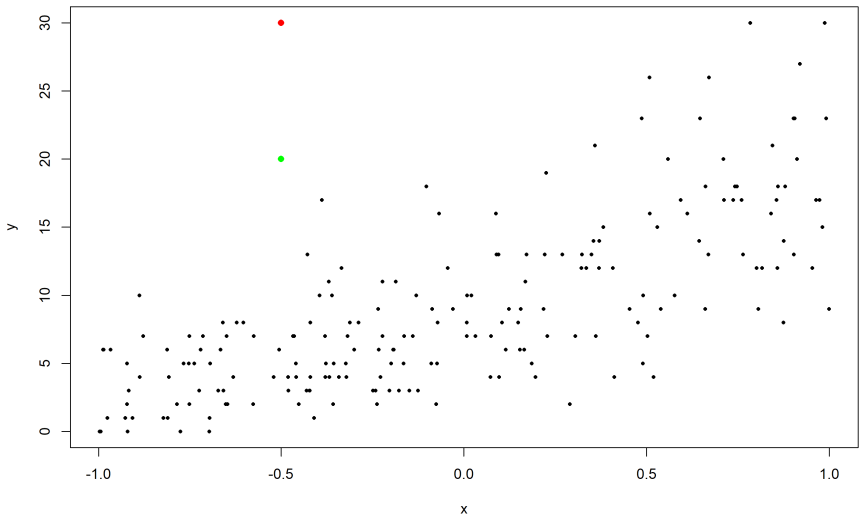


Figure 9: Example of simulated NB data with the single outlier illustrated in red

We observed that the estimates of the regression coefficients of the Poisson-RM and NB-RM exhibit more bias compared to the cNB-RM. This is even more pronounced for the far case. Relatedly, the MSE is lower for the estimates of the regression coefficients of the cNB-RM than for the other two models.

Table 3: Simulation results of sensitivity analysis for added outlier based on 1000 replications.

		Close			Far		
		Poisson-RM	NB-RM	cNB-RM	Poisson-RM	NB-RM	cNB-RM
Bias	β_0	0.0144	0.0147	0.0076	0.0275	0.0278	0.0067
	β_1	-0.0277	-0.0290	-0.0135	-0.0441	-0.0454	-0.0088
MSE	β_0	0.0015	0.0015	0.0015	0.0020	0.0020	0.0014
	β_1	0.0046	0.0045	0.0042	0.0055	0.0053	0.0037

To assess the ability of the cNB-RM to automatically classify whether a point is good or bad using (8), following the approach in [26], we report: (i) the true positive rate (TPR), which measures the proportion of atypical observations that are correctly identified as atypical; and (ii) the false positive rate (FPR), which corresponds to the proportion of typical points incorrectly classified as atypical. The results are reported in Table 4.

Table 4: Classification results of cNB-RM for Close and Far scenarios.

	TPR	FPR
Close	0.9910	0.0304
Far	0.9990	0.0002

For the close case, of the 1000 added atypical observations to the 1000 generated datasets, 991 were classified as atypical, leading to a TPR of 0.991. Similarly, of the $200 \times 1000 = 200000$ typical points, only 6086 were wrongly classified as outliers, leading to an FPR of 0.03043. Additionally, the TPR is greater for the far case compared to the close one, which makes intuitive sense, since the closer the added atypical point is to the generated NB data, the more likely the cNB-RM is to misclassify the outlier as a good point.

5. Real data analysis

The cNB-RM is applied to real-world benchmark datasets, namely the `badhealth` and `azpro` datasets, which are both freely available in the `COUNT` package in `R`. To illustrate the model's viability as an alternative for overdispersed data, we benchmark it to other NB variations, including the linear NB (NB-1), heterogeneous NB (NB-H), and the generalized NB (NB-P), as detailed in [14]. The alternative models are fit using the `VGAM` package, as described in [39].

As mentioned in Section 1, if the observed zeros in the data exceed the distributional assumption of the model, this can also be a cause of overdispersion. Generally, when overdispersion arises as a result of an excess amount of zeros, a suitable strategy is to model the data using either a zero-inflated Poisson (ZIP, [18]) or a zero-inflated negative binomial (ZINB, [38]) model. The model performance is ranked as usual [39] via the Akaike information criterion (AIC; [1]) and the Bayesian information criterion (BIC; [30]). Moreover, the likelihood-ratio (LR) test, which compares nested models, can be used to determine whether the cNB-RM (alternative model) significantly improves upon the NB-RM (null model) since the cNB-RM includes the NB-RM as a special case (see Proposition 1). Under the null hypothesis of no improvement, the test statistic is

$$\text{LR} = -2\left[l(\hat{\boldsymbol{\beta}}, \hat{\alpha}) - l(\hat{\boldsymbol{\beta}}, \hat{\alpha}, \hat{\delta}, \hat{\eta})\right], \tag{11}$$

where $\hat{\boldsymbol{\beta}}, \hat{\alpha}, \hat{\delta}$, and $\hat{\eta}$ are the ML estimates of $\boldsymbol{\beta}, \alpha, \delta$, and η, respectively, and where $l(\hat{\boldsymbol{\beta}}, \hat{\alpha})$ and $l(\hat{\boldsymbol{\beta}}, \hat{\alpha}, \hat{\delta}, \hat{\eta})$ are the maximized log-likelihood values under the NB-RM and cNB-RM, respectively. Using Wilk's theorem, LR can be approximated by a χ^2 random variable with degrees of freedom equal to the difference in the number of parameters between the alternative and null model. This allows us to compute a p-value to assess the significance of improvement.

5.1. Number of visits to a doctor in a year data

In the first application, we refocus our attention on the `badhealth` dataset consisting of $n = 1127$ participants from a comprehensive health study conducted in Germany in 1998. There are three variables: "numvisit", the number of visits to a doctor that year (our Y); "badh", a binary variable with the value 1 for patients who claim to be in bad health, and 0 otherwise; and the patient's age. We wish to quantify the effect of "badh" and age on "numvisit". As evident in Figure 1, "numvisit" has a large number of zero counts present in the data, indicating that the ZIP and ZINB might be suitable models to model the data.

Table 5: Ranking of fitted regression models to `badhealth` data according to AIC and BIC.

Regression model	#par	loglikelihood	AIC	rank	BIC	rank
Poisson-RM	3	-2816.28	5638.55	8	5653.63	8
NB-RM	4	-2233.64	4475.28	3	4495.39	3
cNB-RM	6	-2222.81	**4457.62**	1	**4487.79**	1
NB-1	4	-2247.36	4502.73	6	4522.84	6
NB-H	6	-2225.42	4462.84	2	4493.00	2
NB-P	5	-2233.64	4477.27	4	4502.41	4
ZIP	6	-2549.05	5110.10	7	5140.26	7
ZINB	7	-2231.87	4477.75	5	4512.94	5

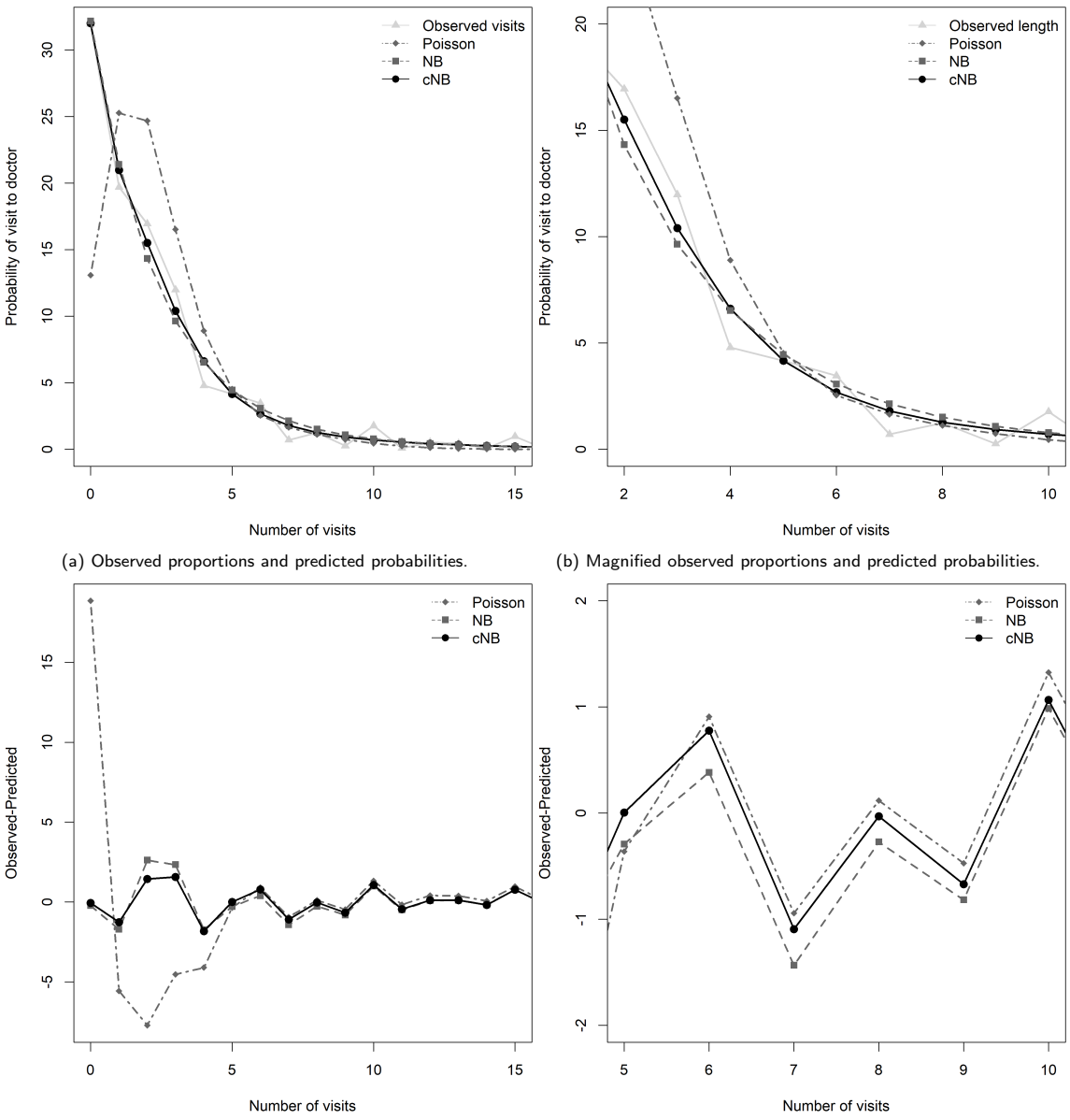
From Table 5, it is apparent that the cNB-RM outperformed all the considered models based on the AIC and BIC. This is further corroborated by the LR test (11) which has a p-value < 0.0001, indicating that cNB-RM is a significant improvement over the NB-RM, regardless of the level of significance chosen. The estimated coefficients for the NB-RM and cNB-RM and the corresponding SEs are reported in Table 6. The SEs for the intercept and "badh" coefficients are similar between the two models but are slightly lower for the age coefficient of the cNB-RM. We note that, since $\hat{\delta} = 0.427$, approximately 42.7% of the observations can be considered as outliers, and have a degree of contamination of 9.309. This is similar to the result of the detection rule in (8), which classified 37.09% of the observations as bad. The observed proportions and predicted probabilities of the Poisson-RM, NB-RM, and cNB-RM are depicted in Figure 10(a), while the difference between the observed proportions and predicted probabilties is presented in Figure 10(c). The close association between the number of visits to a doctor and the predicted number of visits on the basis of the cNB-RM is better observable in the magnified versions in Figures 10(b) and (d). In Figures 10(c) and (d) parts of the lines that are above 0 on the y-axis indicate underprediction of counts while parts of the line below indicate overprediction. Notably, the cNB-RM provides the most accurate predictions among the models considered, as evidenced by the difference between the observed proportions and predicted probabilities being closest to 0.

Table 6: Estimated coefficients and corresponding SEs (in brackets) of NB-RM and cNB-RM to `badhealth` data.

Parameter	NB-RM	cNB-RM
Intercept	0.404 (0.131)	0.549 (0.132)
badh	1.107 (0.112)	1.158 (0.106)
age	0.007 (0.003)	0.003 (0.003)
$\hat{\alpha}$	1.003 (0.070)	0.280 (0.124)
$\hat{\delta}$		0.425 (0.121)
$\hat{\eta}$		9.291 (3.423)

5.2. Heart procedures data

The `azpro` dataset includes records of $n = 3589$ patients entering an Arizona hospital in 1991 to receive one of two standard cardiovascular treatments (PTCA = percutaneous transluminal coronary angioplasty = 0, CABG = coronary artery bypass surgery = 1) called the procedure variable. The other variables are admit (0 = elective, 1 = urgent/emergency), age75 (0 if age $<$ 75 years, otherwise 1), and sex (M = 1, F = 0). The response is the integer-valued length of hospital stay (los), in days. The primary objective is to compare the effectiveness of the



(a) Observed proportions and predicted probabilities.

(b) Magnified observed proportions and predicted probabilities.

(c) Difference between observed proportions and predicted probabilities.

(d) Magnified difference between observed proportions and predicted probabilities.

Figure 10: Observed proportions and predicted probabilities for the number of doctor visits.

two treatments while accounting for the influence of the other variables. That is, whether the difference in stay is statistically significant between the two procedures, controlling for gender, type of admission, and patient age. Determining the probable length of stay is also desirable given patient profiles. Based on the results presented in Table 7, it is clear that the cNB-RM outperformed all the considered alternative models via the AIC and BIC. This is supported by the LR test which has a p-value < 0.0001. The estimated coefficients for the NB-RM and cNB-RM and the corresponding SEs are reported in Table 8, where it is observed that the SEs are smaller for the cNB-RM coefficients. The observed proportions and predicted probabilities of the length of stay, and the difference between them, as predicted by the Poisson-RM, NB-RM, and cNB-RM are illustrated in Figure 11.

Table 7: Ranking of fitted models to azpro data according to the AIC and BIC.

Regression Model	#par	loglikelihood	AIC	rank	BIC	rank
Poisson-RM	5	-11189.90	22389.80	6	22420.72	6
NB-RM	6	-9973.54	19959.09	5	19996.20	5
cNB-RM	8	-9828.82	**19673.64**	1	**19723.13**	1
NB-1	6	-9960.39	19932.80	4	19969.91	4
NB-H	10	-9933.47	19886.93	2	19948.79	2
NB-P	7	-9948.52	19911.03	3	19954.33	3

Table 8: Estimated coefficients and corresponding SEs (in brackets) of NB and cNB regression models fitted to azpro data.

Parameter	NB-RM	cNB-RM
Intercept	1.418 (0.024)	1.374 (0.022)
procedure	0.981 (0.018)	0.972 (0.017)
sex	-0.126 (0.019)	-0.121 (0.017)
admit	0.371 (0.019)	0.352 (0.017)
age75	0.120 (0.020)	0.121 (0.018)
$\hat{\alpha}$	0.160 (0.007)	0.057 (0.008)
$\hat{\delta}$		0.109 (0.022)
$\hat{\eta}$		14.409 (1.967)

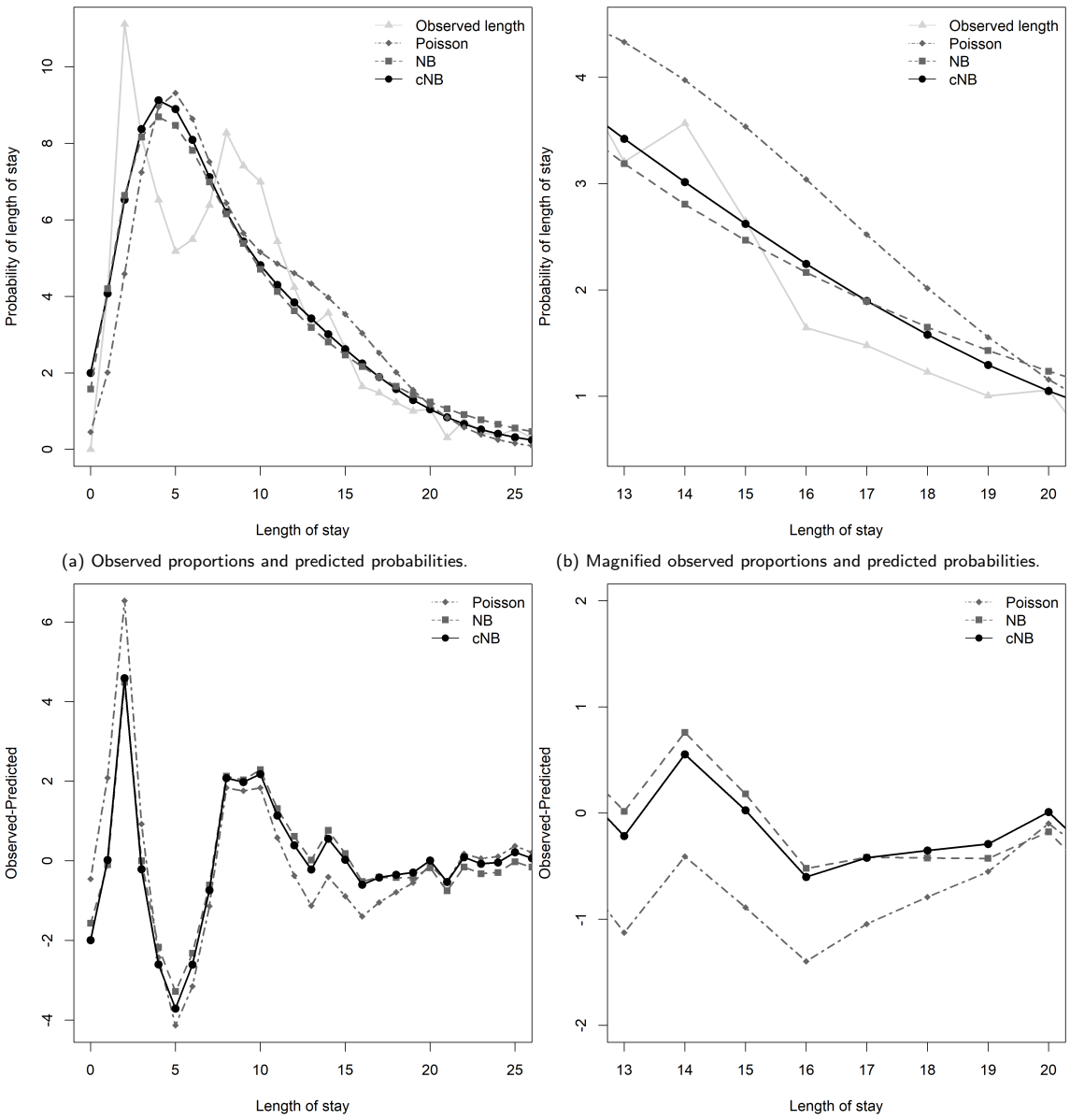
6. Conclusion

In this paper, we introduced a straightforward extension of the negative binomial regression model (NB-RM), termed contaminated NB-RM (cNB-RM). This model not only addresses overdispersion more effectively than the NB-RM but also offers greater flexibility in accommodating the conditional skewness and excess kurtosis observed in real (health) data. Relatedly, real-world data frequently includes mild outliers and extreme values, which significantly impact all the statistical moments largely discussed in this paper, namely mean, variance, skewness, and kurtosis. Advantageously, our proposed model is formulated as a simple mixture of two NB-RMs with the same means but different dispersion parameters, and this formulation not only retains a closed-form expression for the probability mass function of the conditional response variable but also provides the capability, if desired, to identify mild outliers. These outliers can be understood as observations stemming from the contaminant NB-RM, distinct from the regular observations associated with the reference NB-RM. Last but not least, the simple formulation of our proposal involves two additional parameters having practical interpretation, an aspect of fundamental importance not only for statisticians but also for practitioners who use statistical models and want to interpret the output from the considered model. These parameters are the proportion of observations originating from the contaminant NB-RM (potentially considered as mild outliers or extreme values) and the degree of contamination. This degree of contamination roughly quantifies how dispersed the observations in the contaminant NB-RM are compared to those in the regular distribution.

Furthermore, an expectation-maximization (EM) algorithm for maximum likelihood estimation of the parameters of the cNB-RM is proposed. A parameter recovery study is performed to assess the EM algorithm's accuracy in retrieving the true generating parameters. The impact of outliers on parameter estimation and their potential to introduce bias into the regression parameters, which could distort inferences and overestimate the overdispersion parameter of the NB-RM when it compensates for the outliers, are examined through a sensitivity analysis.

Regarding real-world scenarios, we utilized the cNB-RM on two benchmark health datasets and compared its performance with other NB variations, including a zero-inflated NB-RM. In both cases, the cNB-RM demonstrated superior performance over the considered regression models, highlighting its viability as an alternative model for overdispersed count data. While the proposed models are motivated by applications in health, their use is not restricted to this field and other fields may benefit from its use.

Future extensions of the cNB-RM could include allowing the contamination parameters δ and η to be modeled as functions of covariates, similar to the approach often used in zero-inflated regression models. By incorporating covariate-dependent contamination parameters, the model would gain additional flexibility.



(a) Observed proportions and predicted probabilities.

(b) Magnified observed proportions and predicted probabilities.

(c) Difference between observed proportions and predicted probabilities.

(d) Magnified difference between observed proportions and predicted probabilities.

Figure 11: Observed proportions and predicted probabilities for the length of stay.

Acknowledgements

Ferreira and Bekker have been partially supported by: (i) the National Research Foundation (NRF) of South Africa (SA), grant RA201125576565, nr 145681; RA171022270376, Grant No: 119109; and (ii) the DSI-NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS) - grant 2024-033-STA and 2024-034-STA, South Africa. Bekker also acknowledges the support of the National Research Foundation (NRF) of South Africa (SA) ref. SRUG2204203865 nr. 120839. The opinions expressed and conclusions arrived at are those of the authors and are not necessarily to be attributed to the NRF.

Data availability statement

All datasets considered in this paper are freely available in R.

Disclosure statement

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723.

[2] Barnett, V. and T. Lewis (1994). *Outliers in Statistical Data*, Volume 3. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics.

[3] Berk, R. and J. M. MacDonald (2008). Overdispersion and Poisson regression. *Journal of Quantitative Criminology 24*, 269–284.

[4] Berzel, A., G. Z. Heller, and W. Zucchini (2006). Estimating the number of visits to the doctor. *Australian and New Zealand Journal of Statistics 48*(2), 213–224.

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

[6] Cameron, A. C. and P. K. Trivedi (2013). *Regression analysis of count data*. Cambridge University Press.

[7] Davies, L. and U. Gather (1993). The identification of multiple outliers. *Journal of the American Statistical Association 88*(423), 782–792.

[8] Fernandez, G. A. and K. P. Vatcheva (2022). A comparison of statistical methods for modeling count data with an application to hospital length of stay. *BMC Medical Research Methodology 22*(1), 211.

[9] Frome, E. L. and H. Checkoway (1985). Use of Poisson regression models in estimating incidence rates and ratios. *American Journal of Epidemiology 121*(2), 309–323.

[10] Green, J. A. (2021). Too many zeros and/or highly skewed? A tutorial on modelling health behaviour as count data with Poisson and negative binomial regression. *Health Psychology and Behavioral Medicine 9*(1), 436–455.

[11] Greenwood, M. and G. U. Yule (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal statistical society 83*(2), 255–279.

[12] Group, S. (2001). The German Socio-Economic Panel (GSOEP) after more than 15 years: overview. *Proceedings of the 2000 Fourth International Conference of German Socio-Economic Panel Study Users (GSOEP2000), Vierteljahrshefte zur Wirtschaftsforschung 70*(1), 7–14.

[13] Hennig, C. (2002). Fixed point clusters for linear regression: computation and comparison. *Journal of Classification 19*(2), 249.

[14] Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

[15] Ismail, N. and A. A. Jemain (2007). Handling overdispersion with negative binomial and generalized Poisson regression models. In *Casualty Actuarial Society Forum*, Volume 2007, pp. 103–58. Citeseer.

[16] Kenne Pagui, E. C., A. Salvan, and N. Sartori (2022). Improved estimation in negative binomial regression. *Statistics in Medicine 41*(13), 2403–2416.

[17] Klakattawi, H. S., V. Vinciotti, and K. Yu (2018). A simple and adaptive dispersion regression model for count data. *Entropy 20*(2), 142.

[18] Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics 34*(1), 1–14.

[19] Lindén, A. and S. Mäntyniemi (2011). Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology 92*(7), 1414–1421.

[20] Mazza, A. and A. Punzo (2020). Mixtures of multivariate contaminated normal regression models. *Statistical Papers 61*(2), 787–822.

[21] Morris, K., A. Punzo, P. D. McNicholas, and R. P. Browne (2019). Asymmetric clusters and outliers: Mixtures of multivariate contaminated shifted asymmetric Laplace distributions. *Computational Statistics & Data Analysis 132*, 145–166.

[22] Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics 33*(3), 341–365.

[23] Mwalili, S. M., E. Lesaffre, and D. Declerck (2008). The zero-inflated negative binomial regression model with correction for misclassification: An example in caries research. *Statistical Methods in Medical Research 17*(2), 123–139.

[24] Preisser, J. S., K. Das, D. L. Long, and K. Divaris (2016). Marginalized zero-inflated negative binomial regression with application to dental caries. *Statistics in Medicine 35*(10), 1722–1735.

[25] Punzo, A. and L. Bagnato (2021). The multivariate tail-inflated normal distribution and its application in finance. *Journal of Statistical Computation and Simulation 91*(1), 1–36.

[26] Punzo, A. and P. D. McNicholas (2016). Parsimonious mixtures of multivariate contaminated normal distributions. *Biometrical Journal 58*(6), 1506–1537.

[27] Punzo, A. and P. D. McNicholas (2017). Robust clustering in regression analysis via the contaminated Gaussian cluster-weighted model. *Journal of Classification 34*, 249–293.

[28] Punzo, A. and C. Tortora (2021). Multiple scaled contaminated normal distribution and its application in clustering. *Statistical Modelling 21*(4), 332–358.

[29] Ritter, G. (2014). *Robust cluster analysis and variable selection*. CRC Press.

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

[31] Thomas, W. J., K. E. Guire, and G. G. Horvat (1997). Is patient length of stay related to quality of care? *Journal of Healthcare Management 42*(4), 489–507.

[32] Tomarchio, S. D., L. Bagnato, and A. Punzo (2022). Model-based clustering via new parsimonious mixtures of heavy-tailed distributions. *AStA Advances in Statistical Analysis 106*(2), 315–347.

[33] Tomarchio, S. D., M. P. Gallaugher, A. Punzo, and P. D. McNicholas (2022). Mixtures of matrix-variate contaminated normal distributions. *Journal of Computational and Graphical Statistics 31*(2), 413–421.

[34] Tomarchio, S. D., A. Punzo, and L. Bagnato (2020). Two new matrix-variate distributions with application in model-based clustering. *Computational Statistics & Data Analysis 152*, 107050.

[35] Tomarchio, S. D., A. Punzo, J. T. Ferreira, and A. Bekker (2024). A new look at the Dirichlet distribution: Robustness, clustering, and both together. *Journal of Classification*, 1–23.

[36] Tortora, C., B. C. Franczak, L. Bagnato, and A. Punzo (2024). A Laplace-based model with flexible tail behavior. *Computational Statistics and Data Analysis 192*, 107909.

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

[38] Yau, K. K., K. Wang, and A. H. Lee (2003). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal 45*(4), 437–452.

[39] Yee, T. W. (2020). The VGAM package for negative binomial regression. *Australian and New Zealand Journal of Statistics 62*(1), 116–131.

[40] Zhang, Y., V. Melnykov, and I. Melnykov (2023). On model-based clustering of directional data with heavy tails. *Journal of Classification 40*(3), 527–551.

Appendix A. Proofs

Proof Proposition 1: The cNB-D in (4) has the hierarchical representation

$$\begin{aligned} W &\sim \mathcal{B}_{\{1,\eta\}}(\delta) \\ Y|W=w &\sim \mathcal{NB}(\mu, w\alpha), \end{aligned} \tag{A.1}$$

where $\mathcal{B}_{\{1,\eta\}}(\delta)$ denotes a Bernoulli random variable with probability of success δ on the support $\{1,\eta\}$ defined as

$$W = \begin{cases} 1 & \text{with probability } 1-\delta, \\ \eta & \text{with probability } \delta. \end{cases} \tag{A.2}$$

The proofs of (*a*)–(*d*) in Proposition 1 follow:

(*a*) if $\delta \to 0^+$, from (A.2) it follows that $W \xrightarrow{D} 1$ and, therefore, according to (A.1)–(A.2), $Y \xrightarrow{D} \mathcal{NB}(\mu, \alpha)$;

(*b*) if $\eta \to 1^+$, from (A.2) it follows that $W \xrightarrow{D} 1$ and, as before, according to (A.1)–(A.2), $Y \xrightarrow{D} \mathcal{NB}(\mu, \alpha)$;

(*c*) if $\delta \to 0^+$ and $\alpha \to 0^+$, from the proof for (*a*) and from the results given in [11], it follows that $Y \xrightarrow{D} \mathcal{P}\text{ois}(\mu)$;

(*d*) if $\eta \to 1^+$ and $\alpha \to 0^+$, from the proof for (*b*) and, again, as demonstrated in [11], it follows that $Y \xrightarrow{D} \mathcal{P}\text{ois}(\mu)$.