Environmetrics

WILEY

RESEARCH ARTICLE OPEN ACCESS

Modeling Bounded Count Environmental Data Using a Contaminated Beta-Binomial Regression Model

Arnoldus F. Otto[1] | Antonio Punzo[2] | Johannes T. Ferreira[3] | Andriëtte Bekker[1] | Salvatore D. Tomarchio[2] | Cristina Tortora[4]

[1]Department of Statistics, University of Pretoria, Pretoria, South Africa | [2]Department of Economics and Business, University of Catania, Catania, Italy | [3]School of Statistics and Actuarial Science, University of the Witwatersrand, Johannesburg, South Africa | [4]Department of Mathematics and Statistics, San José State University, San Jose, California, USA

Correspondence: Arnoldus F. Otto (arno.otto@up.ac.za)

Received: 18 April 2025 | **Revised:** 27 November 2025 | **Accepted:** 11 December 2025

Keywords: beta-binomial | climate data analysis | contaminated beta-binomial distribution | count data | count data regression modeling | kurtosis | overdispersion

ABSTRACT

Bounded count data are commonly encountered in environmental studies. This paper examines two environmental applications illustrating their relevance. The first investigates the effect of winter malnutrition on mule deer (*Odocoileus hemionus*) fawn mortality. The second application analyzes public perceptions of environmental issues using data from the Eurobarometer 95.1 survey (March–April 2021), which includes a question rating the perceived severity of climate change on a scale from 1 to 10. Together, these studies demonstrate the need for flexible bounded count models in environmental research. In this context, the binomial and beta-binomial (BB) models are widely used for bounded count data, with the BB model offering the advantage of accounting for overdispersion. However, atypical observations in real-world applications may hinder the performance of the BB model and lead to biased or misleading inferences. To address this limitation, we propose the contaminated beta-binomial (cBB) distribution (cBB-D), which introduces an additional BB component to accommodate atypical observations while preserving the mean and variance structure of the BB model. The cBB-D thus captures both overdispersion and contamination effects in bounded count data. To incorporate explanatory variables, we further develop the contaminated BB regression model (cBB-RM), in which none, some, or all cBB parameters may depend on covariates. The proposed models are applied to two environmental datasets, complemented by a sensitivity analysis on simulated data to assess the influence of atypical observations on parameter estimation. The methodology is implemented in the open-source **cBB** package for R, available at https://github.com/arnootto/cBB.

1 | Introduction

In environmental research, data often consist of bounded counts, such as the natural counts of different species or the number of species affected by natural or human-induced events (Paul and Saha 2007; Ryan 2007). Considerable attention has been given to the study of anthropogenic environmental changes, as these can influence both ecosystems (Yee et al. 2008) and economies (Sciandra et al. 2024). The relevance of bounded count data in this field, however, extends well beyond ecological contexts. For instance, the number of citations issued for hazardous emissions (among many other regulatory indicators) can have important implications for environmental monitoring and enforcement (e.g., Konisky and Woods 2010; Spina 2015). Collectively, these

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Environmetrics* published by John Wiley & Sons Ltd.

examples highlight the central role of bounded count data and the importance of developing appropriate statistical models for their analysis in environmental science.

From the perspective of applications to bounded count data, this paper presents two case studies in environmental research that illustrate distinct yet complementary aspects of environmental and climatic phenomena. The first application analyzes animal counts following an environmental event, while the second investigates public perceptions of environmental issues using survey data.

The first dataset investigates the relationship between winter malnutrition and mule deer (*Odocoileus hemionus*) fawn mortality. Variations in environmental conditions can significantly impact species survival and ecosystem dynamics, making such analyses crucial for understanding population processes (Aspinall and Matthews 1994; Muluneh 2021). Mule deer inhabit diverse environments, and their population dynamics are influenced by multiple factors. Particularly notable are abrupt changes, such as population declines following harsh winters, as well as longer-term fluctuations occurring over broader spatial scales. The dataset analyzed here was collected from 1875 radio-collared fawns captured in early winter across Colorado, Idaho, and Montana (USA) between 1981 and 1996 (Unsworth et al. 1999). This study provides an extensive examination of overwinter survival in a species sensitive to environmental variability, offering insights into the ecological effects of changing environmental conditions.

The second application explores public perceptions of climate-related issues, which are relevant for understanding societal attitudes toward environmental policies and practices. Public opinion can shape support for policy measures and influence the adoption of sustainable behaviors (Arikan and Günay 2021). The Eurobarometer is a long-running series of public opinion surveys conducted on behalf of the European Commission and other European Union (EU) institutions since 1973. These surveys examine a range of topics across EU member states, including attitudes toward environmental and climate issues. In the Eurobarometer 95.1 survey, conducted between March and April 2021, one key question posed was:

> How serious a problem do you think climate change is at this moment? Please use a scale from 1 to 10, with '1' meaning it is 'not at all a serious problem' and '10' meaning it is 'an extremely serious problem'.

This question provides a quantitative assessment of the perceived importance of climate change, capturing individual evaluations of its relevance and potential impact. The ordinal response scale, ranging from 1 to 10, allows for a detailed examination of how respondents perceive the extent of the issue. The inclusion of personal information enables further analysis across demographic and socioeconomic groups. Surveys of this kind often include several key variables expressed as bounded counts, which can be analyzed using appropriate statistical models to investigate patterns in environmental attitudes and related behaviors.

Methodologically, as we will detail in Section 2, numerous parametric probability distributions have been proposed to model bounded count data.

The binomial model and its regression extension have long been the default options because they directly address bounded counts. In the binomial setting, a single parameter (i.e., the so-called probability of success, which we will denote by π and define in Section 2) governs the main distributional features: the average number of successes, the variability around that average, and the shape characteristics such as asymmetry and tail weight. This parsimony is attractive, but it also limits adaptability. In particular, when the observed variability is larger than what the binomial model allows — an issue known as overdispersion — the fit can be systematically inadequate.

The beta-binomial (BB) distribution provides a natural remedy by allowing π to vary across observational units. This introduces a dispersion parameter that inflates variability while leaving the mean structure aligned with the binomial model, thereby accommodating heterogeneous populations and unobserved factors. As dispersion increases, observations near the extremes of the support become more likely, reflecting real-world scenarios in which all-or-nothing outcomes occur more frequently than the binomial model predicts. Despite its flexibility, the BB distribution still relies on only two parameters, which can be insufficient to capture higher-order features such as skewness and excess kurtosis. When many extreme observations are present, this limitation may lead to understated standard errors (SEs) and an overstatement of statistical significance, with the risk of misleading inferences.

Motivated by these considerations, we introduce the contaminated BB (cBB) distribution (cBB-D). This model is formulated as a two-component mixture of BB distributions, where one component represents the typical (or regular) observations (the reference BB-D) and the other, with the same mean but an increased dispersion, which accounts for the excess of extreme observations (the contaminant distribution). The underlying approach mirrors that of other studies that utilize contaminated distributions to address varying data supports; see, for example, Punzo (2019), Punzo and Bagnato (2021, 2025), Otto, Ferreira, Tomarchio, et al. (2025), and Tomarchio et al. (2025). For a detailed discussion of the reference distribution concept, which we assume to be the BB-D in this case, refer to Davies and Gather (1993).

The proposed cBB-D has the advantage of a closed-form probability mass function (PMF) and interpretable parameters. As highlighted by Ley et al. (2021), Otto, Ferreira, Bekker, et al. (2025), and Wagener et al. (2024), it is crucial for parameters to have meaningful interpretations to draw valid inferences about the underlying population from which the data originate. In detail, in addition to the standard BB-D parameters, the cBB-D introduces two new parameters: one representing the proportion of observations from the contaminant BB-D and another indicating the degree of contamination. The additional parameters enable the cBB-D to capture empirical skewness and excess kurtosis more effectively, and, by extension, address the excess of extreme observations compared to the reference BB-D. Furthermore, a cBB

regression model (cBB-RM) is developed by incorporating the cBB-D into a regression framework, allowing for the inclusion of available and relevant covariates. Unlike traditional RMs, we do not limit the regression to the parameter π only, but, considering convenient link functions, we extend the regression to all the parameters of the cBB-D, and, to further increase the flexibility of the method, we also allow for different covariates on each parameter.

The paper is organized as follows. Section 2 provides the background for the methodology under consideration and serves as the foundation for our proposal. Section 3 presents the cBB distribution and its regression extension. Section 4 outlines an expectation–maximization (EM) algorithm for maximum likelihood (ML) estimation, together with initialization strategies and convergence criteria. Section 5 reports a simulation study assessing parameter recovery and a sensitivity analysis examining the impact of extreme observations. Section 6 illustrates the methodology on two environmental datasets—the mule deer survival data and the Eurobarometer survey—demonstrating the practical advantages of the cBB model for overdispersed bounded counts. Section 7 concludes and sketches directions for future research. The main distributional results of practical interest are collected in Appendix A, and a practitioner's guide to implementing the methodology is provided in Appendix B.

2 | Background

Let Y denote the bounded count variable of interest, taking values in $\{0, 1, \ldots, m\}$, where $m \geq 1$ is known in advance. It is important to note that, in a statistical framework, Y is often interpreted as the number of successes out of m trials. For the discussion that follows, it is useful to keep in mind both interpretations of Y: as a bounded count in general, and as a count of successes.

As discussed in Section 1, handling sample data for Y poses a nuanced challenge. Although it may be tempting to reduce such data to proportions, it is generally preferable to model the raw counts directly rather than converting them to proportions before analysis. This approach appropriately weights larger trials more heavily than smaller ones, instead of treating all proportions equally. For example, a proportion of 0.5 could arise from very different scenarios, such as a single success out of two attempts or four successes out of eight. Focusing solely on proportions obscures the context underlying the data and limits the ability to capture the true variability inherent in count-based observations (Martin et al. 2020).

2.1 | The Binomial Distribution

The one-parameter binomial distribution (B-D) and its extension, the binomial regression model (B-RM), have traditionally been the standard choices as they directly model the raw bounded counts. The B-D is defined by the following PMF:

$$f_{\mathrm{B}_m}(y;\pi) = \binom{m}{y}\pi^y(1-\pi)^{m-y}, \quad y = 0, 1, \ldots, m, \tag{1}$$

where $\pi \in (0, 1)$ represents the probability of success in each trial. The notation reflects the fact that in the considered context, π is the only unknown parameter, whereas m is an inherent characteristic of the phenomenon under study. If Y has the PMF in (1), we denote it as $Y \sim \mathcal{B}_m(\pi)$. The moments and characteristics of practical interest of $Y \sim \mathcal{B}_m(\pi)$, namely mean, variance, skewness, and excess kurtosis, are:

$$\begin{aligned}
\mathrm{E}_{\mathrm{B}_m}(Y;\pi) &= m\pi, \\
\mathrm{Var}_{\mathrm{B}_m}(Y;\pi) &= m\pi(1-\pi), \qquad (2)\\
\mathrm{Skew}_{\mathrm{B}_m}(Y;\pi) &= \frac{1-2\pi}{\sqrt{m\pi(1-\pi)}}, \qquad (3)
\end{aligned}$$

and

$$\mathrm{ExKurt}_{\mathrm{B}_m}(Y;\pi) = \frac{1-6\pi(1-\pi)}{m\pi(1-\pi)}. \tag{4}$$

All these characteristics are governed by the single parameter π. The variance in (2) lies in $(0, m/4]$, attaining its maximum at $\pi = 0.5$. The skewness in (3) depends on π, being positive for $\pi < 0.5$, negative for $\pi > 0.5$, and zero when $\pi = 0.5$; it is always bounded within the interval $(-1/\sqrt{m\pi(1-\pi)}, 1/\sqrt{m\pi(1-\pi)})$. As m increases, the skewness decreases in magnitude, approaching zero. The excess kurtosis in (4) reaches its minimum at $\pi = 0.5$, taking the value $-2/m$, and diverges to infinity as π approaches 0 or 1. Consequently, as m increases, the range of possible negative excess kurtosis values narrows. Although the characteristics may vary, their dependence solely on π restricts the binomial model's ability to capture the empirical behavior of bounded counts accurately. In particular, it is well known that the binomial model is inadequate in the presence of overdispersion, where the observed variance exceeds the binomial variance for a given mean.

2.2 | The Beta-Binomial Distribution

The BB distribution (BB-D) has emerged as a natural extension of the B-D to address the latter limitation, and plays the same role for the B-D as the negative binomial distribution does for the Poisson distribution. Its flexibility makes it a powerful tool in various fields, including epidemiology (Arostegui et al. 2010) and microbiology (Martin et al. 2020), where the underlying success probabilities (i.e., π) may vary or be uncertain.

The mean-parameterized or, more appropriately, the π-parametrized BB-D has the following PMF:

$$f_{\mathrm{BB}_m}(y;\pi,\sigma) = \binom{m}{y}\frac{\mathrm{B}\left(y+\frac{\pi}{\sigma}, m-y+\frac{1-\pi}{\sigma}\right)}{\mathrm{B}\left(\frac{\pi}{\sigma}, \frac{1-\pi}{\sigma}\right)}, \quad y = 0, 1, \ldots, m, \tag{5}$$

where $\pi \in (0, 1)$ is the analogue of the binomial probability of success parameter, $\sigma > 0$ is the dispersion parameter, and $\mathrm{B}(\cdot,\cdot)$ denotes the beta function. If Y has the PMF in (5), then we simply write $Y \sim \mathcal{BB}_m(\pi, \sigma)$. The expected value and variance of $Y \sim \mathcal{BB}_m(\pi, \sigma)$ are

$$\mathrm{E}_{\mathrm{BB}_m}(Y;\pi) = m\pi, \tag{6}$$

as for the B-D, and

$$\begin{aligned}\mathrm{Var}_{\mathrm{BB}_m}(Y;\pi,\sigma) &= m\pi(1-\pi)\Big[1+(m-1)\frac{\sigma}{1+\sigma}\Big]\\ &= \mathrm{Var}_{\mathrm{B}_m}(Y;\pi)\Big[1+(m-1)\frac{\sigma}{1+\sigma}\Big].\end{aligned} \tag{7}$$

Consequently, the mean in (6) is solely determined by π, whereas the variance in (7) is influenced by both π and σ. The parameterization of the PMF in (5) in terms of π and σ offers enhanced statistical interpretability and, because of (6), allows the use of the BB-D in a regression (toward the mean) context. Since the term $\frac{\sigma}{1+\sigma}$ in (7) lies within $(0,1)$, the term on squared brackets on the right-hand side of (7) can assume values in $(1,m)$ and, as such, it acts as an inflation factor that allows the BB-variance in (7) to span from $\mathrm{Var}_{\mathrm{B}_m}(Y;\pi)$ to $m\mathrm{Var}_{\mathrm{B}_m}(Y;\pi)$, thereby accommodating varying degrees of binomial overdispersion. This is why one can interpret σ as a measure of the overdispersion. Another aspect which is important to emphasize is that, as σ increases in $\mathcal{BB}_m(\pi,\sigma)$, the variance in (7) also increases, making observations at the extremes of the support $\{0,1,\ldots,m\}$ (which we will frequently refer to as extreme observations in this paper) more likely compared to the binomial distribution $\mathcal{B}_m(\pi)$. As a limiting case, in the scenario of maximum variance ($\sigma\to\infty$), the BB-D tends to a two-point distribution assuming values 0 and m with probabilities $1-\pi$ and π, respectively. Some researchers favor the reparameterization in terms of π and $\gamma=\frac{\sigma}{1+\sigma}$; see, for example, Bayes et al. (2024). However, since σ and γ share a one-to-one correspondence, they convey equivalent information about the BB-D. Finally, skewness and excess kurtosis for $Y\sim\mathcal{BB}_m(\pi,\sigma)$ are given by

$$\begin{aligned}\mathrm{Skew}_{\mathrm{BB}_m}(Y;\pi,\sigma) &= \frac{(1-2\pi)(2m\sigma+1)}{(2\sigma+1)\sqrt{\frac{m\pi(1-\pi)(m\sigma+1)}{\sigma+1}}}\\ &= \frac{(1-2\pi)(2m\sigma+1)}{(2\sigma+1)\sqrt{\mathrm{Var}_{\mathrm{BB}_m}(Y;\pi,\sigma)}},\end{aligned} \tag{8}$$

and

$$\mathrm{ExKurt}_{\mathrm{BB}_m}(Y;\pi,\sigma) = \frac{6\pi(1-\pi)(m(6\sigma+5)\sigma(m\sigma+1)+\sigma+1)-(\sigma+1)(\sigma(6m(m\sigma+1)-1)+1)}{\pi(1-\pi)m(2\sigma+1)(3\sigma+1)(m\sigma+1)}. \tag{9}$$

The skewness and excess kurtosis of the $\mathcal{BB}_m(\pi,\sigma)$ are governed by both π and σ. The skewness in (8) is always bounded within the interval

$$\left(-\frac{(2m\sigma+1)}{(2\sigma+1)\sqrt{\frac{(1-\pi)m\pi(m\sigma+1)}{\sigma+1}}},\ \frac{(2m\sigma+1)}{(2\sigma+1)\sqrt{\frac{(1-\pi)m\pi(m\sigma+1)}{\sigma+1}}}\right).$$

As for the B-D case, it is zero when $\pi=0.5$, positive for $\pi<0.5$, and negative for $\pi>0.5$. Increasing σ leads to an increase in the magnitude of the skewness, while an increase in m decreases it. The excess kurtosis in (9) reaches its minimum $-\frac{2(3m\sigma(m\sigma+1)+\sigma+1)}{m(1+3\sigma)(1+m\sigma)}$ at $\pi=0.5$, and diverges to infinity as π approaches 0^+ or 1^-. The excess kurtosis generally increases as σ increases. However, this is not the case when the parameter π lies within the interval

$$\pi\in\Big(\frac{1}{2}-c,\frac{1}{2}+c\Big),$$

where

$$c=\frac{\sqrt{3}}{6}\sqrt{\frac{(2\sigma_1+1)(2\sigma_2+1)(3m(\sigma_2(m\sigma_1+\sigma_1+1)+\sigma_1)+2)}{6m\sigma_2^2(m\sigma_1+\sigma_1+1)+\sigma_2(m\sigma_1+\sigma_1+1)(m(6\sigma_1+5)+6)+\sigma_1(m(6\sigma_1+5)+6)+4}}.$$

Here, σ_1 and σ_2 are two distinct positive values of σ, such that $\sigma_2>\sigma_1>0$. Within this interval for π, the excess kurtosis decreases as σ increases—that is, the excess kurtosis at σ_2 is smaller than at σ_1. For example, if $m=10$, $\sigma_1=0.1$ and $\sigma_2=1$, the excess kurtosis for σ_2 is larger than that for σ_1 when $\pi\in(0,0.223)$ and $\pi\in(0.777,1)$, smaller when $\pi\in(0.223,0.777)$, and equal when $\pi=0.223$ or $\pi=0.777$.

As noted by Bayes et al. (2024), despite the ability of the BB-D to capture binomial overdispersion, it is limited by its two parameters, π and σ, which do not offer sufficient flexibility to fully capture higher-order characteristics such as skewness and excess kurtosis (refer to (8) and (9)). The latter is particularly important as it may indicate an excess of extreme observations relative to the fitted BB-D. Consequently, the BB-D may struggle to model empirical skewness and kurtosis accurately, as it cannot fully adjust the distribution's shape beyond the mean and variance. From an inferential perspective, fitting the BB-D in the presence of extreme observations can lead to underestimation of SEs and overstatement of the significance of estimates, resulting in potentially misleading inferences.

3 | Methodological Proposals

Motivated by the considerations above, we introduce in this section the cBB-D (Section 3.1) and the cBB-RM (Section 3.2).

3.1 | The Contaminated Beta-Binomial Distribution

The PMF of the proposed cBB-D is

$$f_{\mathrm{cBB}_m}(y;\pi,\sigma,\delta,\eta) = (1-\delta)\underbrace{f_{\mathrm{BB}_m}(y;\pi,\sigma)}_{\text{reference}} + \delta\underbrace{f_{\mathrm{BB}_m}(y;\pi,\eta\sigma)}_{\text{contaminant}}, \tag{10}$$

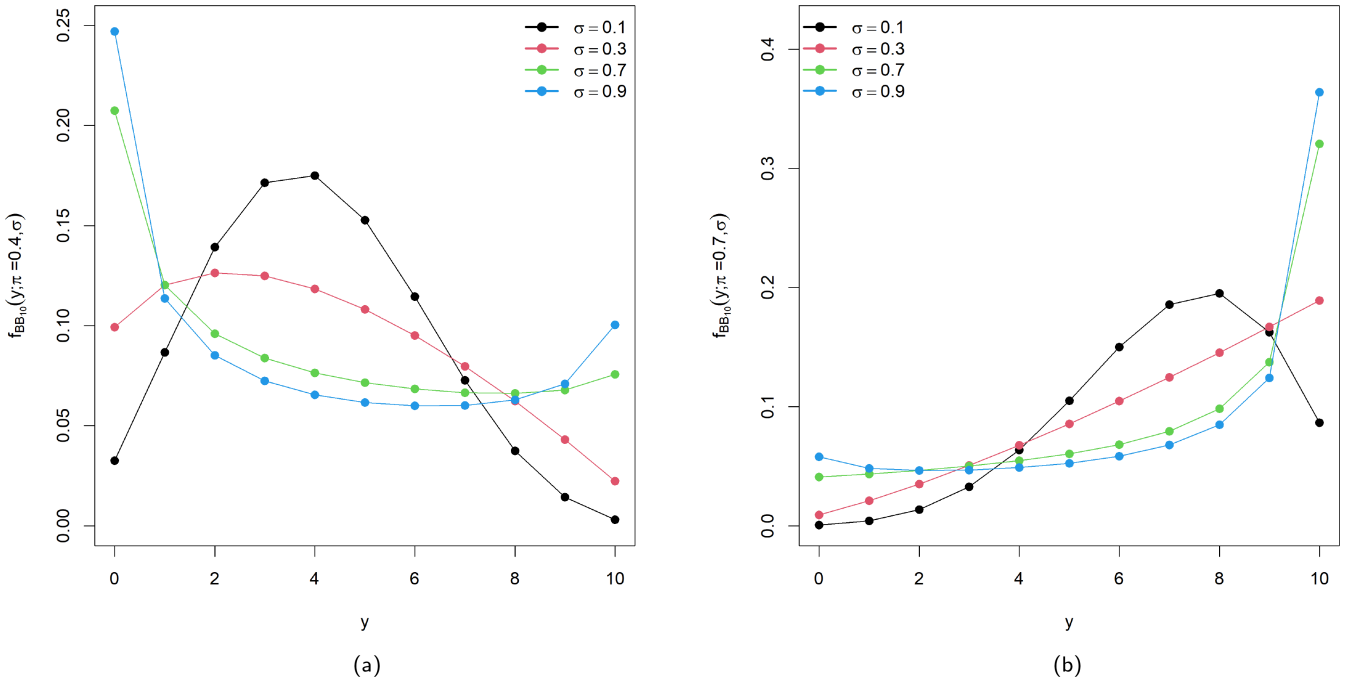


FIGURE 1 | Plots of the BB-D (5) for different values of σ when $m = 10$. (a) $\pi = 0.4$, (b) $\pi = 0.7$.

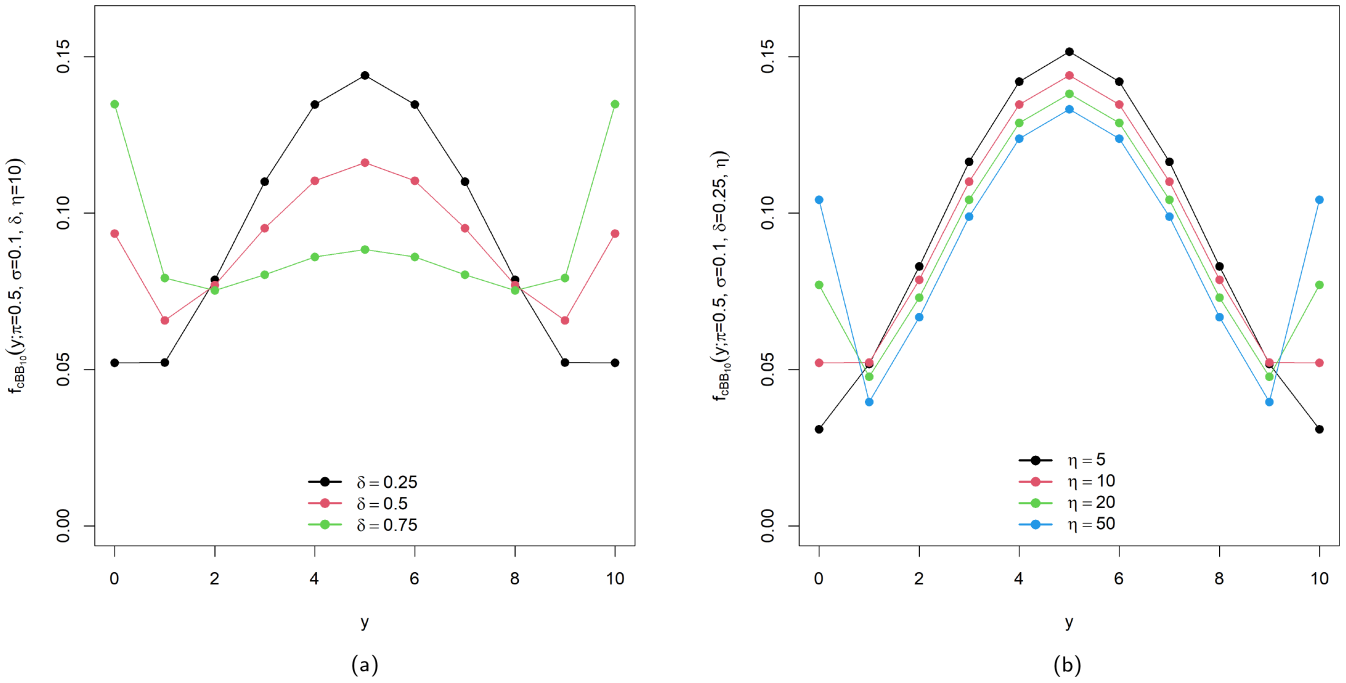


FIGURE 2 | Plots of the cBB-D (10) with $m = 10$, $\pi = 0.5$, and $\sigma = 0.1$ for different values of δ (when $\eta = 10$) and η (when $\delta = 0.25$). (a) Different values of δ, (b) Different values of η.

where $\pi \in (0,1)$, $\sigma > 0$, $\delta \in (0,1)$, and $\eta > 1$. If Y has the PMF given in (10), we will simply write $Y \sim c\mathcal{BB}_m(\pi, \sigma, \delta, \eta)$. Due to the bimodal nature of the BB-D—which can be bell-shaped, U-shaped, J-shaped, or reverse-J-shaped depending on its parameter values—the cBB-D may also be bimodal, or even trimodal. This flexibility allows the cBB-D to model W-shaped data, where data is clustered at both tails like the U-shape of the BB-D, while retaining an additional central mode (Gallop et al. 2013, Keller-Ressel 2022, and Korkmaz 2020). The bimodal nature of the BB-D is illustrated in Figure 1 for varying values of σ, while examples of the cBB-D are illustrated in Figure 2 for different choices of δ and η.

The moments, or shape characteristics, of practical interest of $Y \sim c\mathcal{BB}_m(\pi, \sigma, \delta, \eta)$ are derived in Appendix A and are:

$$\mathrm{E}_{\mathrm{cBB}_m}(Y; \pi) = m\pi, \tag{11}$$

$$\begin{aligned}
&\mathrm{Var}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta)\\
&\quad=(1-\delta)\mathrm{Var}_{\mathrm{BB}_m}(Y;\pi,\sigma)+\delta\mathrm{Var}_{\mathrm{BB}_m}(Y;\pi,\eta\sigma)\\
&\quad=\frac{m\pi(1-\pi)[(1-\delta)(1+m\sigma)(1+\eta\sigma)+\delta(1+m\eta\sigma)(1+\sigma)]}{(1+\sigma)(1+\eta\sigma)},
\end{aligned}\tag{12}$$

$$\begin{aligned}
&\mathrm{Skew}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta)\\
&\quad=\frac{1-\delta}{\left(\mathrm{Var}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta)\right)^{\frac{3}{2}}}\\
&\qquad\times\left(\frac{m\pi(1-\pi)(1-2\pi)(1+m\sigma)(1+2m\sigma)}{(1+\sigma)(1+2\sigma)}\right)\\
&\qquad+\frac{\delta}{\left(\mathrm{Var}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta)\right)^{\frac{3}{2}}}\\
&\qquad\times\left(\frac{m\pi(1-\pi)(1-2\pi)(1+m\sigma\eta)(1+2m\sigma\eta)}{(1+\sigma\eta)(1+2\sigma\eta)}\right),
\end{aligned}\tag{13}$$

$$\begin{aligned}
\mathrm{ExKurt}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta)=-3+\frac{m\pi(1-\pi)}{\left[\mathrm{Var}_{\mathrm{cBB}}(Y;\pi,\sigma,\delta,\eta)\right]^2}\\
\times\left(\frac{(1-\delta)(m\sigma+1)(6(3(\pi-1)\pi+1)m^2\sigma^2+3m\sigma(2-(\pi-1)\pi(m-6))-\ 3(\pi-1)\pi(m-2)-\sigma+1)}{(\sigma+1)(2\sigma+1)(3\sigma+1)}\right.\\
\left.+\frac{\delta(\eta m\sigma+1)(-\eta\sigma+6\eta^2(3(\pi-1)\pi+1)m^2\sigma^2+3\eta m\sigma(2-(\pi-1)\pi(m-6))-\ 3(\pi-1)\pi(m-2)+1)}{(\eta\sigma+1)(2\eta\sigma+1)(3\eta\sigma+1)}\right).
\end{aligned}\tag{14}$$

The variance ranges between $\left(0,\frac{m[(1-\delta)(1+m\sigma)(1+\eta\sigma)+\delta(1+m\eta\sigma)(1+\sigma)]}{4(1+\sigma)(1+\eta\sigma)}\right]$, reaching a maximum at $\pi=0.5$ and tending to 0 as $\pi\to0^+$ or $\pi\to1^-$. The variance increases with σ and, since $\delta\in(0,1)$ and $\eta>1$, the variance of the cBB-D is always greater than that of the BB-D. Moreover, the variance increases as δ and η increase, as illustrated in Figure 3 for different values of δ and η. Similarly, the skewness of the cBB-D is consistently larger in magnitude than the BB-D counterparts, as illustrated in Figure 4. As expected, Figure 4a and b demonstrate that the skewness of the distribution exhibits rotational symmetry around $\pi=0.5$. The distribution exhibits positive skewness when $\pi<0.5$ and negative skewness when $\pi>0.5$. Moreover, the magnitude of the skewness increases monotonically as π deviates further from $\pi=0.5$. The kurtosis is at a minimum for $\pi=0.5$, with the effect of δ and η illustrated in Figure 5. We observe that increasing either δ or η leads to a rise in excess kurtosis for some values of π, while for other values, it causes a decrease: mirroring the effect of increasing σ in the BB-D case. We also note that when $\delta\to0^+$ the kurtosis approaches that of the reference BB-D, as illustrated for $\delta=0.05$ in Figure 5a and c and is consistent with Proposition 1 (a). Since the cBB-D moments depend on four parameters, rather than just π and σ as in the BB-D, the inclusion of δ and η enhances the model's flexibility. This added flexibility allows the cBB-D to better accommodate possible extreme observations and effectively address overdispersion observed in the BB-D. Moreover, the moments of the BB-D emerge as limiting cases of the cBB-D, reinforcing its ability to generalize and extend the BB-D.

In Proposition 1, we explore some limiting cases of the cBB-D.

Proposition 1. *Let $Y\sim c\mathcal{BB}_m(\pi,\sigma,\delta,\eta)$, then:*

a. if $\delta\to0^+$, then $Y\xrightarrow{D}\mathcal{BB}_m(\pi,\sigma)$;

b. if $\eta\to1^+$, then $Y\xrightarrow{D}\mathcal{BB}_m(\pi,\sigma)$;

c. if $\delta\to0^+$ and $\sigma\to0^+$, then $Y\xrightarrow{D}\mathcal{B}_m(\pi)$;

d. if $\eta\to1^+$ and $\sigma\to0^+$, then $Y\xrightarrow{D}\mathcal{B}_m(\pi)$;

where $\xrightarrow{D}$ denotes convergence in distribution.

Proof. See Appendix A. □

3.2 | The Contaminated Beta-Binomial Regression Model

In traditional RMs, the focus is typically placed on modeling only the mean, with the assumption that other parameters, like the dispersion or contamination parameters in our case, remain fixed. This can be limiting, especially in situations where, for example, the dispersion in the data may be influenced by covariates and make the assumption of a constant σ inappropriate. Moreover, we allow for different covariates for each parameter: this offers greater flexibility, which enables the model to account for the distinct factors that affect various aspects of the cBB distribution: the mean, dispersion, proportion of contamination, and degree of contamination.

Let $\boldsymbol{x}$, $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{z}$ represent possible values of the covariates $\boldsymbol{X}$, $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{Z}$, which have dimensions p, q, r, and s, respectively. These covariates are used to model the parameters π, σ, δ, and η, respectively, of the cBB-D. The cBB-RM is then specified through the following link functions:

$$\begin{aligned}
g_1(\pi(\boldsymbol{x};\boldsymbol{\beta}))&=\mathrm{logit}(\pi(\boldsymbol{x};\boldsymbol{\beta}))=\widetilde{\boldsymbol{x}}'\boldsymbol{\beta},\\
g_2(\sigma(\boldsymbol{u};\boldsymbol{\alpha}))&=\log(\sigma(\boldsymbol{u};\boldsymbol{\alpha}))=\widetilde{\boldsymbol{u}}'\boldsymbol{\alpha},\\
g_3(\delta(\boldsymbol{v};\boldsymbol{\gamma}))&=\mathrm{logit}(\delta(\boldsymbol{v};\boldsymbol{\gamma}))=\widetilde{\boldsymbol{v}}'\boldsymbol{\gamma},\\
g_4(\eta(\boldsymbol{z};\boldsymbol{\lambda}))&=\log(\eta(\boldsymbol{z};\boldsymbol{\lambda})-1)=\widetilde{\boldsymbol{z}}'\boldsymbol{\lambda},
\end{aligned}$$

where $\boldsymbol{\beta}=(\beta_0,\beta_1,\ldots,\beta_p)'$, $\boldsymbol{\alpha}=(\alpha_0,\alpha_1,\ldots,\alpha_q)'$, $\boldsymbol{\gamma}=(\gamma_0,\gamma_1,\ldots,\gamma_r)'$, and $\boldsymbol{\lambda}=(\lambda_0,\lambda_1,\ldots,\lambda_s)$ are vectors of unknown regression coefficients, $\widetilde{\boldsymbol{x}}=(1,\boldsymbol{x}')'$, $\widetilde{\boldsymbol{u}}=(1,\boldsymbol{u}')'$, $\widetilde{\boldsymbol{v}}=(1,\boldsymbol{v}')'$, and $\widetilde{\boldsymbol{z}}=(1,\boldsymbol{z}')'$ account for the intercept, and log represents the natural logarithm. Naturally, the considered link functions, even if the most commonly used, are only examples of possible

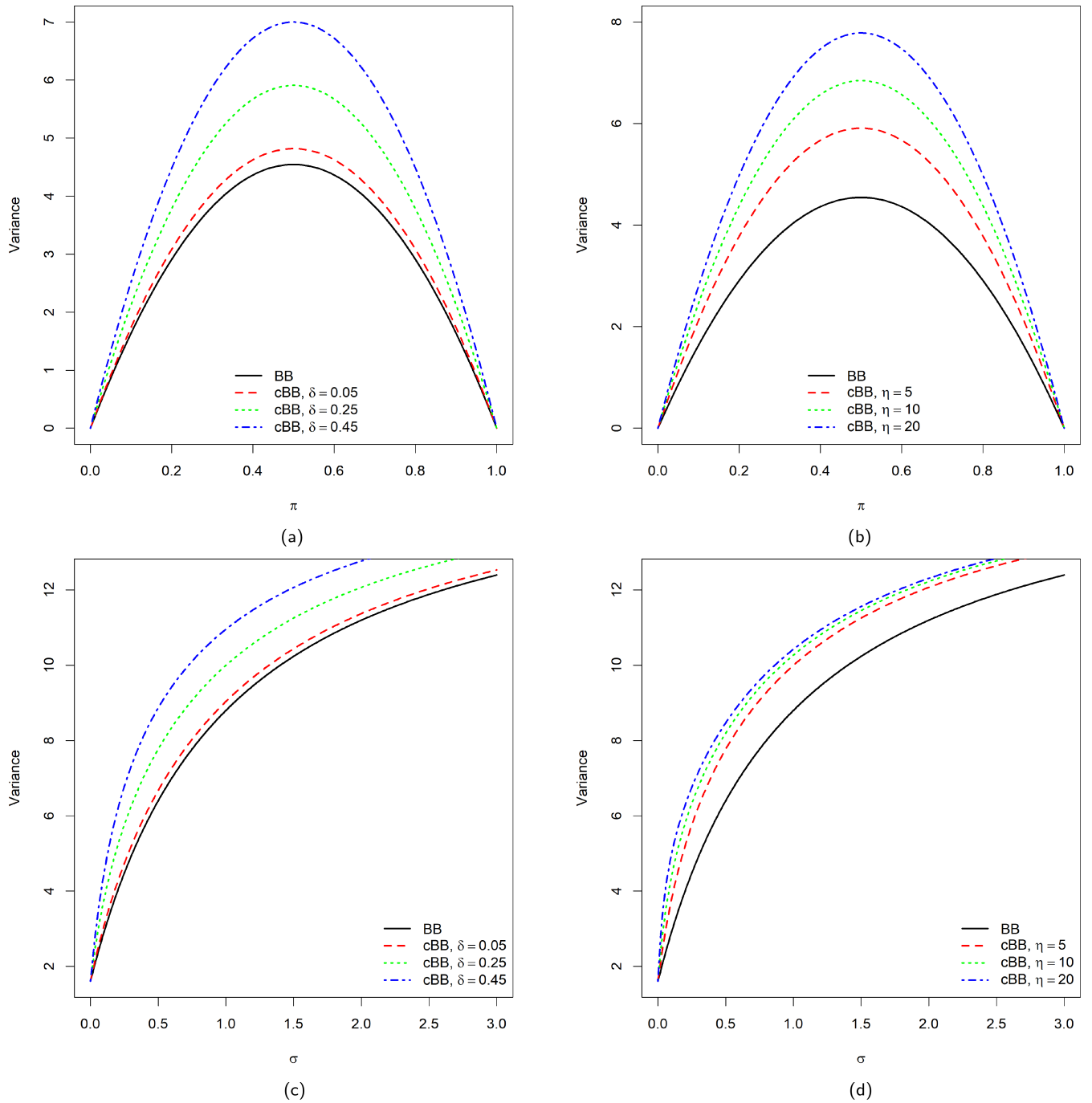


FIGURE 3 | Examples illustrating higher variance (12) of the cBB-D with $m = 10$, $\pi = 0.2$, and $\sigma = 0.1$ compared to the BB-D variance (7) for $\pi \in (0, 1)$ and increasing values of σ for varying values of δ (when $\eta = 5$) and η (when $\delta = 0.25$). (a) Different values of δ. (b) Different values of η. (c) Different values of δ. (d) Different values of η.

functions that can be considered. The inverse of the considered link functions leads to

$$\pi(\boldsymbol{x};\boldsymbol{\beta}) = g_1^{-1}(\widetilde{\boldsymbol{x}}'\boldsymbol{\beta}) = \frac{e^{\widetilde{\boldsymbol{x}}'\boldsymbol{\beta}}}{1+e^{\widetilde{\boldsymbol{x}}'\boldsymbol{\beta}}},$$
$$\sigma(\boldsymbol{u};\boldsymbol{\alpha}) = g_2^{-1}(\widetilde{\boldsymbol{u}}'\boldsymbol{\alpha}) = e^{\widetilde{\boldsymbol{u}}'\boldsymbol{\alpha}},$$
$$\delta(\boldsymbol{v};\boldsymbol{\gamma}) = g_3^{-1}(\widetilde{\boldsymbol{v}}'\boldsymbol{\gamma}) = \frac{e^{\widetilde{\boldsymbol{v}}'\boldsymbol{\gamma}}}{1+e^{\widetilde{\boldsymbol{v}}'\boldsymbol{\gamma}}},$$
$$\eta(\boldsymbol{z};\boldsymbol{\lambda}) = g_4^{-1}(\widetilde{\boldsymbol{z}}'\boldsymbol{\lambda}) = e^{\widetilde{\boldsymbol{z}}'\boldsymbol{\lambda}} + 1.$$

The conditional distribution of Y according to the cBB-RM can also be written as

$$Y|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{U} = \boldsymbol{u}, \boldsymbol{V} = \boldsymbol{v}, \boldsymbol{Z} = \boldsymbol{z} \sim c\mathcal{BB}_m(\pi(\boldsymbol{x};\boldsymbol{\beta}), \sigma(\boldsymbol{u};\boldsymbol{\alpha}), \delta(\boldsymbol{v};\boldsymbol{\gamma}), \eta(\boldsymbol{z};\boldsymbol{\lambda})), \quad (15)$$

with m, which is allowed to vary across sample observations. An advantage of model (15) is that, given the estimates of $\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}$, and $\boldsymbol{\lambda}$, say $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}$, and $\hat{\boldsymbol{\lambda}}$, it is possible to determine whether a data point $(\boldsymbol{x}_i, \boldsymbol{u}_i, \boldsymbol{v}_i, \boldsymbol{z}_i, y_i)$ is an extreme observation or not, with respect to the reference BB-RM, via the a posteriori probability

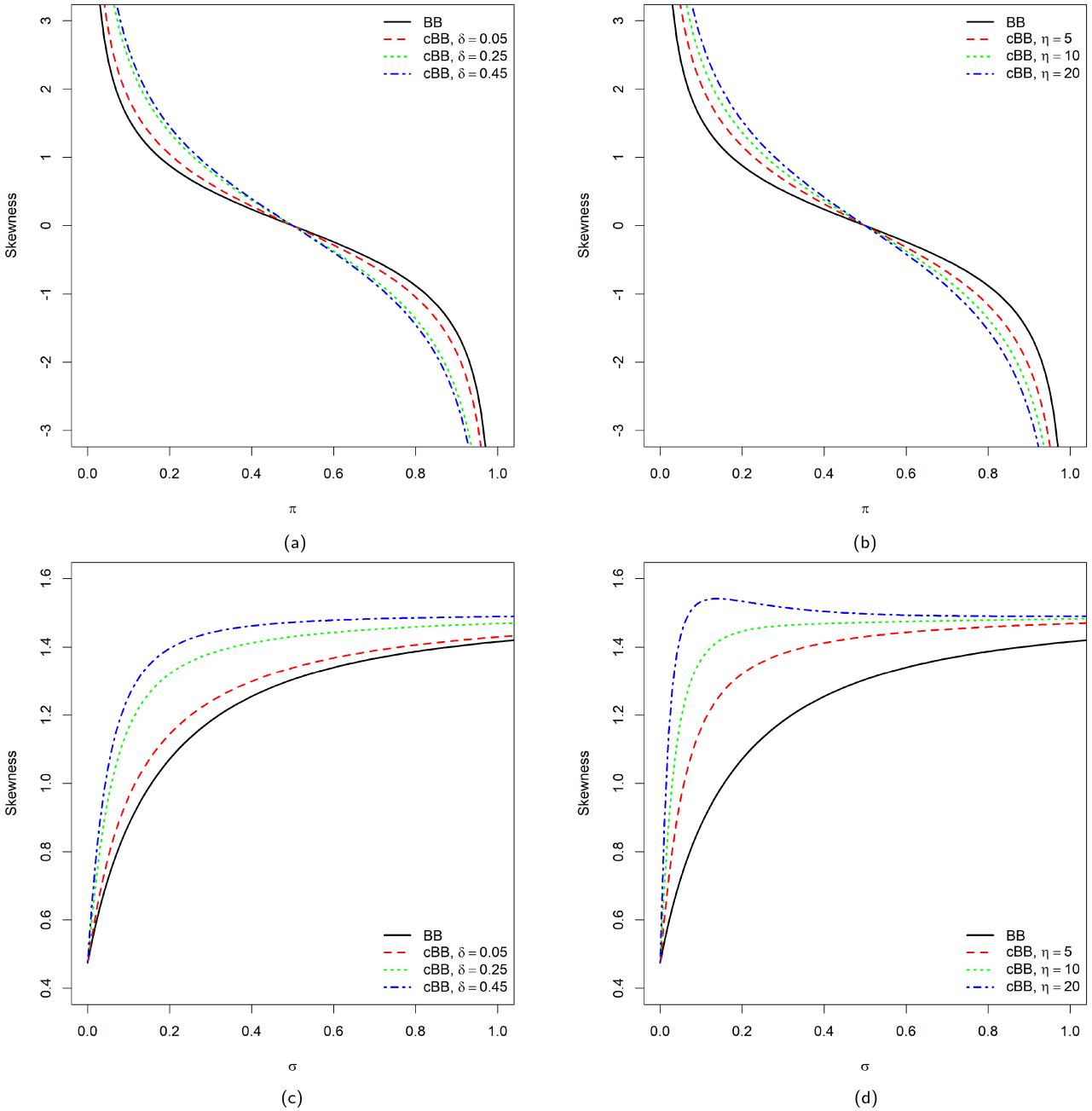


FIGURE 4 | Examples illustrating flexible skewness of the cBB-D (13) with $m = 10$, $\pi = 0.2$, and $\sigma = 0.1$ compared to the BB-D skewness (8) for $\pi \in (0, 1)$ and increasing values of σ for varying values of δ (when $\eta = 5$) and η (when $\delta = 0.25$).

$$P\big((\boldsymbol{x}_i, \boldsymbol{u}_i, \boldsymbol{v}_i, \boldsymbol{z}_i, y_i) \text{ comes from the reference BB-RM} | \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\lambda}}\big) = \frac{(1 - \delta(\boldsymbol{v}; \hat{\boldsymbol{\gamma}}))f_{\text{BB}}(y_i; \pi(\boldsymbol{x}_i; \hat{\boldsymbol{\beta}}), \sigma(\boldsymbol{v}_i; \hat{\boldsymbol{\alpha}}))}{f_{\text{cBB}}(y_i; \pi(\boldsymbol{x}_i; \hat{\boldsymbol{\beta}}), \sigma(\boldsymbol{u}_i; \hat{\boldsymbol{\alpha}}), \delta(\boldsymbol{v}_i; \hat{\boldsymbol{\gamma}}), \eta(\boldsymbol{z}_i; \hat{\boldsymbol{\lambda}}))}. \tag{16}$$

It is natural to consider $(\boldsymbol{x}_i, \boldsymbol{u}_i, \boldsymbol{v}_i, \boldsymbol{z}_i, y_i)$ as an observation from the reference BB-RM if the probability in (16) is greater than 0.5, and an extreme observation otherwise.

4 | Maximum Likelihood Estimation: An EM Algorithm

In this section, we present an EM algorithm for ML estimation of the parameters of the cBB-RM (Section 4.1), followed by a discussion on the initialization strategy and convergence criteria used (Section 4.2).

4.1 | An EM Algorithm

Let $(\boldsymbol{x}_1', \boldsymbol{u}_1', \boldsymbol{v}_1', \boldsymbol{z}_1', y_1), \ldots, (\boldsymbol{x}_n', \boldsymbol{u}_n', \boldsymbol{v}_n', \boldsymbol{z}_n', y_n)$ be the observed sample from (15), with $m_1, \ldots, m_n$ denoting the maximum possible counts for each unit. 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 $(\boldsymbol{x}_i', \boldsymbol{u}_i', \boldsymbol{v}_i', \boldsymbol{z}_i', y_i)$ comes from the reference or contaminant BB-RM. To denote the source of incompleteness, we use an indicator vector $\boldsymbol{w} =$

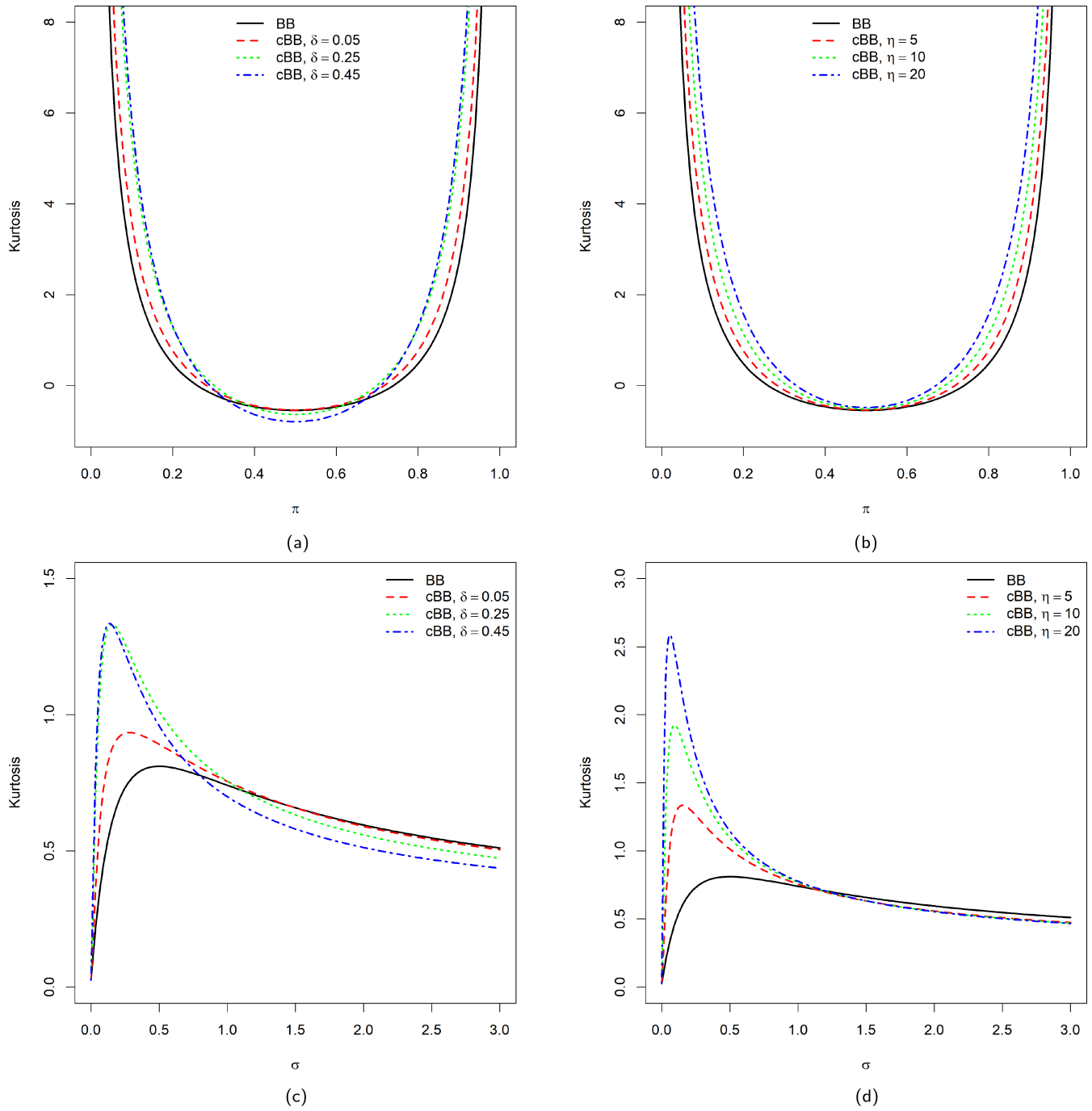


FIGURE 5 | Examples illustrating larger excess kurtosis of the cBB-D (14) with $m = 10$, $\pi = 0.2$, and $\sigma = 0.1$ compared to the BB-D excess kurtosis (9) for $\pi \in (0, 1)$ and increasing values of σ for varying values of δ (when $\eta = 5$) and η (when $\delta = 0.25$). (a) Different values of δ. (b) Different values of η. (c) Different values of δ. (d) Different values of η.

$(w_1, \ldots, w_n)$ so that $w_i = 1$ if $(\boldsymbol{x}_i', \boldsymbol{u}_i', \boldsymbol{v}_i', \boldsymbol{z}_i', y_i)$ comes from the contaminant BB-RM and $w_i = 0$ otherwise. The complete-data are thus given by $(\boldsymbol{x}_1', \boldsymbol{u}_1', \boldsymbol{v}_1', \boldsymbol{z}_1', y_1, w_1), \ldots, (\boldsymbol{x}_n', \boldsymbol{u}_n', \boldsymbol{v}_n', \boldsymbol{z}_n', y_n, w_n)$ and, from (10), the complete-data likelihood function can be written as

$$L_c(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \lambda) = \prod_{i=1}^{n} \left[(1 - \delta(\boldsymbol{v}_i; \boldsymbol{\gamma})) f_{\mathrm{BB}_{m_i}}(y_i; \pi(\boldsymbol{x}_i; \boldsymbol{\beta}), \sigma(\boldsymbol{u}_i; \boldsymbol{\alpha})) \right]^{1-w_i} \times \left[\delta(\boldsymbol{v}_i; \boldsymbol{\gamma}) f_{\mathrm{BB}_{m_i}}(y_i; \pi(\boldsymbol{x}_i; \boldsymbol{\beta}), \eta(\boldsymbol{z}_i; \lambda)\sigma(\boldsymbol{u}_i; \boldsymbol{\alpha})) \right]^{wi}$$

$$= \prod_{i=1}^{n} \left[\frac{(1 - \delta(\boldsymbol{v}_i; \boldsymbol{\gamma}))m_i!}{y_i!(m_i - y_i)!} \frac{\mathrm{B}\left(y_i + \frac{\pi(\boldsymbol{x}_i;\boldsymbol{\beta})}{\sigma(\boldsymbol{u}_i;\boldsymbol{\alpha})}, m_i - y_i + \frac{1-\pi(\boldsymbol{x}_i;\boldsymbol{\beta})}{\sigma(\boldsymbol{u}_i;\boldsymbol{\alpha})}\right)}{\mathrm{B}\left(\frac{\pi(\boldsymbol{x}_i;\boldsymbol{\beta})}{\sigma(\boldsymbol{u}_i;\boldsymbol{\alpha})}, \frac{1-\pi(\boldsymbol{x}_i;\boldsymbol{\beta})}{\sigma(\boldsymbol{u}_i;\boldsymbol{\alpha})}\right)} \right]^{1-w_i} \times \left[\frac{\delta(\boldsymbol{v}_i; \boldsymbol{\gamma})m_i!}{y_i!(m_i - y_i)!} \frac{\mathrm{B}\left(y_i + \frac{\pi(\boldsymbol{x}_i;\boldsymbol{\beta})}{\eta(\boldsymbol{z}_i;\lambda)\sigma(\boldsymbol{u}_i;\boldsymbol{\alpha})}, m_i - y_i + \frac{1-\pi(\boldsymbol{x}_i;\boldsymbol{\beta})}{\eta(\boldsymbol{z}_i;\lambda)\sigma(\boldsymbol{u}_i;\boldsymbol{\alpha})}\right)}{\mathrm{B}\left(\frac{\pi(\boldsymbol{x}_i;\boldsymbol{\beta})}{\eta(\boldsymbol{z}_i;\lambda)\sigma(\boldsymbol{u}_i;\boldsymbol{\alpha})}, \frac{1-\pi(\boldsymbol{x}_i;\boldsymbol{\beta})}{\eta(\boldsymbol{z}_i;\lambda)\sigma(\boldsymbol{u}_i;\boldsymbol{\alpha})}\right)} \right]^{w_i}.$$

The complete log-likelihood function then follows as

$$l_c(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \lambda) = l_{c_1}(\boldsymbol{\gamma}) + l_{c_2}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \lambda) \tag{17}$$

where

$$l_{c_1}(\boldsymbol{\gamma}) = \sum_{i=1}^{n}(1-w_i)\log(1-\delta(\boldsymbol{v}_i;\boldsymbol{\gamma})) + w_i \log \delta(\boldsymbol{v}_i;\boldsymbol{\gamma})$$

and

$$\begin{aligned}
&l_{c_2}(\boldsymbol{\beta},\boldsymbol{\alpha},\lambda)\\
&\quad= \sum_{i=1}^{n}\log\left(\frac{m_i!}{y_i!(m_i-y_i)!}\right)\\
&\qquad+(1-w_i)\left[\log\mathrm{B}\left(y_i+\frac{\pi(\boldsymbol{x}_i;\boldsymbol{\beta})}{\sigma(\boldsymbol{u}_i;\boldsymbol{\alpha})}, m_i-y_i+\frac{1-\pi(\boldsymbol{x}_i;\boldsymbol{\beta})}{\sigma(\boldsymbol{u}_i;\boldsymbol{\alpha})}\right)\right.\\
&\qquad\left.-\log\mathrm{B}\left(\frac{\pi(\boldsymbol{x}_i;\boldsymbol{\beta})}{\sigma(\boldsymbol{u}_i;\boldsymbol{\alpha})}, \frac{1-\pi(\boldsymbol{x}_i;\boldsymbol{\beta})}{\sigma(\boldsymbol{u}_i;\boldsymbol{\alpha})}\right)\right]\\
&\qquad+w_i\left[\log\mathrm{B}\left(y_i+\frac{\pi(\boldsymbol{x}_i;\boldsymbol{\beta})}{\eta(\boldsymbol{z}_i;\lambda)\sigma(\boldsymbol{u}_i;\boldsymbol{\alpha})}, m_i-y_i+\frac{1-\pi(\boldsymbol{x}_i;\boldsymbol{\beta})}{\eta(\boldsymbol{z}_i;\lambda)\sigma(\boldsymbol{u}_i;\boldsymbol{\alpha})}\right)\right.\\
&\qquad\left.-\log\mathrm{B}\left(\frac{\pi(\boldsymbol{x}_i;\boldsymbol{\beta})}{\eta(\boldsymbol{z}_i;\lambda)\sigma(\boldsymbol{u}_i;\boldsymbol{\alpha})}, \frac{1-\pi(\boldsymbol{x}_i;\boldsymbol{\beta})}{\eta(\boldsymbol{z}_i;\lambda)\sigma(\boldsymbol{u}_i;\boldsymbol{\alpha})}\right)\right].
\end{aligned}$$

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

4.1.1 | E-Step

In the E-step, the conditional expectation of the complete-data log-likelihood function is computed as

$$\begin{aligned}
&Q\left(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma},\lambda|\boldsymbol{\beta}^{(k)},\boldsymbol{\alpha}^{(k)},\boldsymbol{\gamma}^{(k)},\lambda^{(k)}\right)\\
&\quad= Q_1\left(\boldsymbol{\gamma}|\boldsymbol{\beta}^{(k)},\boldsymbol{\alpha}^{(k)},\boldsymbol{\gamma}^{(k)},\lambda^{(k)}\right)+Q_2\left(\boldsymbol{\beta},\boldsymbol{\alpha},\lambda|\boldsymbol{\beta}^{(k)},\boldsymbol{\alpha}^{(k)},\boldsymbol{\gamma}^{(k)},\lambda^{(k)}\right)
\end{aligned}$$

for the $(k+1)$-th iteration, which is in the same order as (17). $Q\left(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma},\lambda|\boldsymbol{\beta}^{(k)},\boldsymbol{\alpha}^{(k)},\boldsymbol{\gamma}^{(k)},\lambda^{(k)}\right)$ is obtained by substituting w_i in (17) by the expected a posteriori probability for a point to be an extreme value

$$\begin{aligned}
&\mathrm{E}\left(W_i|y_i,\boldsymbol{x}_i,\boldsymbol{u}_i,\boldsymbol{v}_i,\boldsymbol{z}_i;\boldsymbol{\beta}^{(k)},\boldsymbol{\alpha}^{(k)},\boldsymbol{\gamma}^{(k)},\lambda^{(k)}\right)\\
&\quad=\frac{\delta^{(k)}f_{\mathrm{BB}_{m_i}}\left(y_i;\pi\left(\boldsymbol{x}_i;\boldsymbol{\beta}^{(k)}\right),\eta\left(\boldsymbol{z}_i;\lambda^{(k)}\right)\sigma\left(\boldsymbol{u}_i;\boldsymbol{\alpha}^{(k)}\right)\right)}{f_{\mathrm{cBB}_{m_i}}\left(y_i;\pi\left(\boldsymbol{x}_i;\boldsymbol{\beta}^{(k)}\right),\sigma\left(\boldsymbol{u}_i;\boldsymbol{\alpha}^{(k)}\right),\delta\left(\boldsymbol{v}_i;\boldsymbol{\gamma}^{(k)}\right),\eta\left(\boldsymbol{z}_i;\lambda^{(k)}\right)\right)}:=w_i^{(k)}.
\end{aligned}$$

4.1.2 | M-Step

An update $\boldsymbol{\gamma}^{(k+1)}$ for $\boldsymbol{\gamma}$ is calculated by independently maximizing

$$\begin{aligned}
&Q_1\left(\boldsymbol{\gamma}|\boldsymbol{\beta}^{(k)},\boldsymbol{\alpha}^{(k)},\boldsymbol{\gamma}^{(k)},\lambda^{(k)}\right)\\
&\quad=\sum_{i=1}^{n}\left\{\left(1-w_i^{(k)}\right)\log\left(1-\delta\left(\boldsymbol{v}_i;\boldsymbol{\gamma}^{(k)}\right)\right)+w_i^{(k)}\log\delta\left(\boldsymbol{v}_i;\boldsymbol{\gamma}^{(k)}\right)\right\}
\end{aligned}$$

with respect to $\boldsymbol{\gamma}$ and subjects to constraints on this parameter. Updates for $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and λ are obtained by maximizing

$$\begin{aligned}
&Q_2\left(\boldsymbol{\beta},\boldsymbol{\alpha},\lambda|\boldsymbol{\beta}^{(k)},\boldsymbol{\alpha}^{(k)},\boldsymbol{\gamma}^{(k)},\lambda^{(k)}\right)\\
&\quad=\sum_{i=1}^{n}\left(1-w_i^{(k)}\right)\left[\log\mathrm{B}\left(y_i+\frac{\pi\left(\boldsymbol{x}_i;\boldsymbol{\beta}^{(k)}\right)}{\sigma\left(\boldsymbol{u}_i;\boldsymbol{\alpha}^{(k)}\right)}, m_i-y_i+\frac{1-\pi\left(\boldsymbol{x}_i;\boldsymbol{\beta}^{(k)}\right)}{\sigma\left(\boldsymbol{u}_i;\boldsymbol{\alpha}^{(k)}\right)}\right)\right.\\
&\qquad\left.-\log\mathrm{B}\left(\frac{\pi\left(\boldsymbol{x}_i;\boldsymbol{\beta}^{(k)}\right)}{\sigma\left(\boldsymbol{u}_i;\boldsymbol{\alpha}^{(k)}\right)}, \frac{1-\pi\left(\boldsymbol{x}_i;\boldsymbol{\beta}^{(k)}\right)}{\sigma\left(\boldsymbol{u}_i;\boldsymbol{\alpha}^{(k)}\right)}\right)\right]\\
&\qquad+w_i^{(k)}\left[\log\mathrm{B}\left(y_i+\frac{\pi\left(\boldsymbol{x}_i;\boldsymbol{\beta}^{(k)}\right)}{\eta\left(\boldsymbol{z}_i;\lambda^{(k)}\right)\sigma\left(\boldsymbol{u}_i;\boldsymbol{\alpha}^{(k)}\right)}, m_i-y_i+\frac{1-\pi\left(\boldsymbol{x}_i;\boldsymbol{\beta}^{(k)}\right)}{\eta\left(\boldsymbol{z}_i;\lambda^{(k)}\right)\sigma\left(\boldsymbol{u}_i;\boldsymbol{\alpha}^{(k)}\right)}\right)\right.\\
&\qquad\left.-\log\mathrm{B}\left(\frac{\pi\left(\boldsymbol{x}_i;\boldsymbol{\beta}^{(k)}\right)}{\eta\left(\boldsymbol{z}_i;\lambda^{(k)}\right)\sigma\left(\boldsymbol{u}_i;\boldsymbol{\alpha}^{(k)}\right)}, \frac{1-\pi\left(\boldsymbol{x}_i;\boldsymbol{\beta}^{(k)}\right)}{\eta\left(\boldsymbol{z}_i;\lambda^{(k)}\right)\sigma\left(\boldsymbol{u}_i;\boldsymbol{\alpha}^{(k)}\right)}\right)\right].
\end{aligned}$$

This can be achieved in R software (R Core Team, (R Core Team 2020)) using the `optim()` function included in the **stats** package. The Nelder-Mead or BFGS algorithms, which are used for unconstrained optimization, can be passed to `optim()` via the `method` argument. The algorithm iterates between the E-step and M-step until convergence and is elaborated on below.

4.2 | Initialization and Convergence

The initial values are an essential element in EM-based algorithms and can significantly influence the accuracy and reliability of the model estimates; therefore, their choice represents a vital aspect of the estimation process. Should the initial values be selected inadequately, the algorithm might settle at a local maximum rather than achieving the global maximum. Furthermore, if the initial values significantly diverge from the actual values, the algorithm may experience slow convergence.

We recommend applying a standard BB-RM utilizing the same predictors as those used for cBB-RM in the data analysis. After that, the computed coefficients can be used as starting points for cBB-RM fitting. For δ and η, we suggest choosing them such that the cBB-RM tends to the BB-RM, that is, $\delta^{(0)}\to 1^-$ and $\eta^{(0)}\to 1^+$ (see Proposition 1).

Several convergence criteria can be applied to determine whether or not the EM algorithm has converged. A prevalent approach involves monitoring the variation in the log-likelihood function across successive iterations. If the change falls below a predetermined threshold, the algorithm can be considered to be converged, that is $l(\boldsymbol{\beta}^{(k+1)},\boldsymbol{\alpha}^{(k+1)},\boldsymbol{\gamma}^{(k+1)},\lambda^{(k+1)})-l(\boldsymbol{\beta}^{(k)},\boldsymbol{\alpha}^{(k)},\boldsymbol{\gamma}^{(k)},\lambda^{(k)})<\epsilon$. Due to the possibility of flat likelihoods, we chose a stopping criteria of $\epsilon=1\times10^{-10}$ or 1000 iterations.

5 | Simulation Study: Sensitivity Analysis

In this study, we performed a sensitivity analysis to examine how extreme observations affect the estimates of the B-RM, BB-RM, and cBB-RM. We generate 1000 datasets of sizes $n=500$ from the B-D, with $m=10$, an intercept of $\beta_0=-3$, and a continuous covariate generated by a uniform distribution over the interval $(0,1)$ with slope $\beta_1=7$. The generated data is then augmented with extreme values using one of the following schemes:

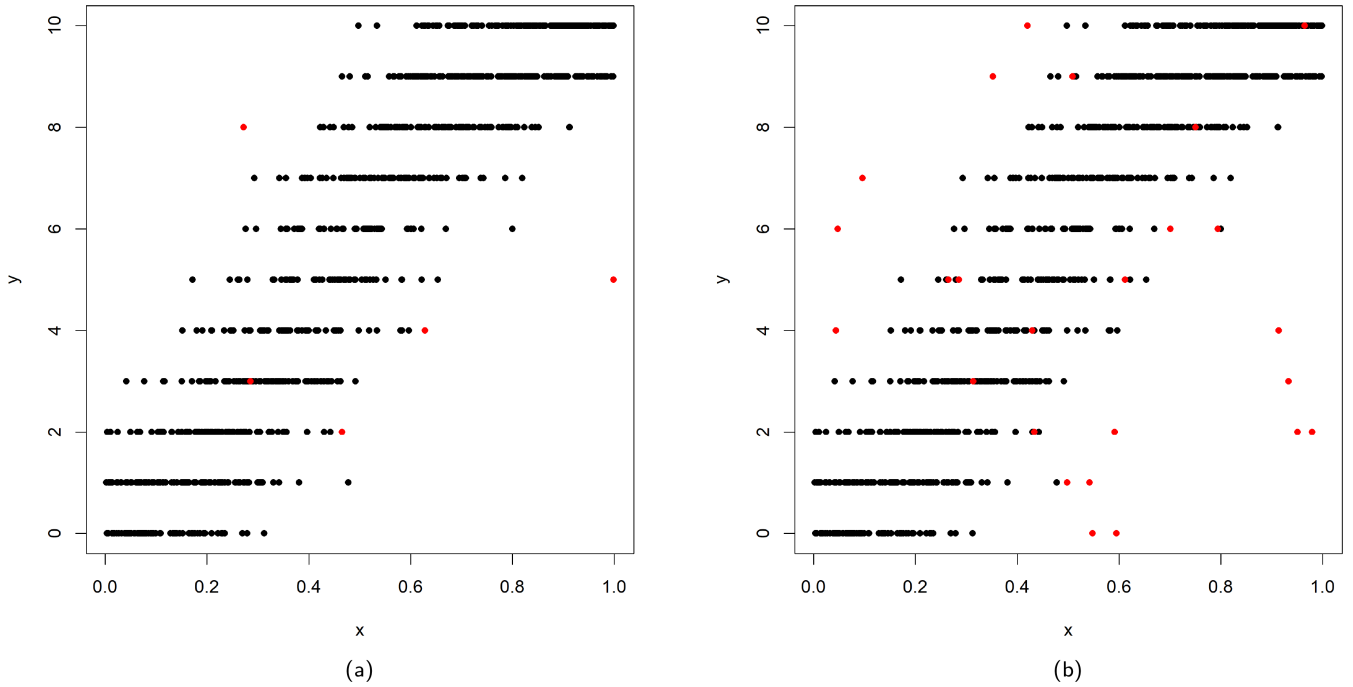


FIGURE 6 | Examples of simulated binomial data with a different percentage of artificially added extreme observations (in red). (a) 1%, (b) 5%.

1. 1%of the generated Y-values are randomly substituted by data generated from a discrete uniform distribution over the set $\{0, 1, \ldots, 10\}$.
2. 5%of the generated Y-values are randomly substituted by data generated from a discrete uniform distribution over the set $\{0, 1, \ldots, 10\}$.

Examples of the aforementioned schemes are illustrated in Figure 6, where the artificially modified observations are highlighted in red. The B-RM, BB-RM, and cBB-RM models are then fitted to the data to assess the impact of the red observations on the estimated regression coefficients. The bias and mean squared error (MSE), reported in Table 1, are computed as

$$\text{Bias}(\hat{\beta}_h) = \left(\frac{1}{1000} \sum_{r=1}^{1000} \hat{\beta}_{hr} \right) - \beta_h$$

and

$$\text{MSE}(\hat{\beta}_h) = \frac{1}{1000} \sum_{r=1}^{1000} \left(\hat{\beta}_{hr} - \beta_h \right)^2,$$

where $\hat{\beta}_{hr}$ is the estimate of β_h, $h = 0, 1$, obtained at the rth replication $r = 1, \ldots, 1000$.

The sensitivity analysis reveals that the presence of extreme observations significantly affects the estimates of the B-RM as reflected by the bias and MSE. Although the BB-RM performs better than the B-RM, it is also sensitive to the extremities. This is even more pronounced as the percentage of extreme observations increases from 1% to 5%, as expected. The cBB-RM consistently demonstrates the lowest bias and MSE for both regression coefficients, indicating that the cBB-RM is more reliable for handling data prone to contamination.

To evaluate the cBB-RM's capability to detect extreme observations, we consider the true positive rate (TPR), measuring the proportion of extreme observations correctly detected, and the false positive rate (FPR), measuring the proportion of good observations incorrectly detected as extreme. The results are reported in Table 2.

As shown in Table 2, the TPR increases slightly from 0.3768 under the 1% scenario to 0.4374 under the 5% scenario. The lack of convergence of the TPR toward one does not necessarily indicate an error. Because of the way the extreme observations are incorporated into the data, some of them may resemble typical observations, leading the cBB-RM to classify them as typical. Meanwhile, the FPR remains consistently low, demonstrating that the cBB-RM rarely misclassifies good observations as extreme.

6 | Real Data Applications

In this section, we investigate the behavior of the cBB-D and cBB-RM by applying them to real-world datasets, namely survival data of mule deer due to winter malnutrition (Section 6.1) and the Eurobarometer 95.1 survey (Section 6.2). To illustrate the model's viability as an alternative for overdispersed bounded counts, we benchmark it to a flexible BB generalization known as the beta-2-binomial (B2B) model (Bayes et al. 2024). Model performance is ranked via the Akaike information criterion (AIC; Akaike 1974),

$$\text{AIC} = 2k - 2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\lambda}), \tag{18}$$

the Bayesian information criterion (BIC; Schwarz 1978),

$$\text{BIC} = \log(n)k - 2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\lambda}),$$

TABLE 1 | Sensitivity analysis to examine how the artificially added anomalous values affect the estimates of the RMs.

		1%			5%		
		B-RM	**BB-RM**	**cBB-RM**	**B-RM**	**BB-RM**	**cBB-RM**
Bias	$\hat{\beta}_0$	0.0678	0.0606	0.0375	0.3138	0.2826	0.1707
	$\hat{\beta}_1$	−0.1636	−0.1438	−0.0886	−0.7529	−0.6653	−0.3996
MSE	$\hat{\beta}_0$	0.0136	0.0123	0.0096	0.1105	0.0904	0.0386
	$\hat{\beta}_1$	0.0643	0.0561	0.0414	0.6204	0.4882	0.1986

TABLE 2 | Classification results of cBB-RM for 1% and 5% scenarios.

	TPR	**FPR**
1% scenario	0.3768	0.0102
5% scenario	0.4374	0.0078

and the Hannan-Quinn information criterion (HQIC; Hannan and Quinn 1979)

$$\text{HQIC} = 2\log(\log(n))k - 2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\lambda}),$$

where k is the number of parameters, n is the number of observations, $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}$, and $\hat{\lambda}$ are the ML estimates of $\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}$ and λ, respectively, and where $l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\lambda})$ is the maximized log-likelihood value for the cBB-RM.

Moreover, we use the likelihood-ratio (LR) test, which evaluates nested models, to assess whether the cBB-RM (alternative model) provides a significant improvement over the BB-RM (null model), as the latter is a special case of the former. Under the null hypothesis of no improvement, the test statistic is

$$\text{LR} = -2[l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}) - l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\lambda})], \tag{19}$$

where $l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$ is the maximized log-likelihood value for the BB-RM and $l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\lambda})$ is as previously defined. Based on Wilk's theorem, the LR statistic approximately follows a χ^2 distribution with degrees of freedom equal to the number of parameters between the alternative and null models. This enables the calculation of a p-value to determine the significance of the improvement.

After using the EM algorithm outlined in Section 4.1 to estimate the model parameters, we obtain the variance-covariance matrix of the parameter estimates by inverting the negative Hessian matrix, which is computed using the `optim()` function in R. The SEs of the cBB-RM parameter estimates are then determined by taking the square roots of the diagonal elements of this matrix.

6.1 | Mule Deer Mortality

The mortality rates of mule deer (*Odocoileus hemionus*) fawns due to winter malnutrition were analyzed using data collected from radio-collared fawns captured during early winter in Colorado, Idaho, and Montana in the United States of America between 1981 and 1996 (Unsworth et al. 1999). The study consisted of 26 separate observations encompassing a total of $n = 1875$ radio-collared mule deer collected over the years in the three states. This dataset represents a comprehensive, long-term study of overwinter survival in a species that is highly sensitive to environmental conditions.

As a first step in the analysis, we fit an intercept-only cBB-RM to deaths (Y), representing the number of mule deer that died to winter malnutrition out of a total of m_i radio-collared fawns, for $i = 1, \ldots, 26$. This is equivalent to fitting the cBB-D, but with varying m. Table 3 presents the results of comparing the cBB-D to the B-D, BB-D, and B2B-D. The models are separately ranked via their AIC, BIC, and HQIC values. The cBB-RM performed the best, but it is also worth noting that the binomial distribution is not enough for this dataset, regardless of the considered measure. The LR test further supports that the cBB-RM is an improvement over the BB-RM with a p-value of 0.007. As for the cBB-RM, it is worth noting that the estimate of the contamination proportion is $\hat{\delta} = 0.603$, showing a large excess of extreme values with respect to the reference cBB-RM for the regular counts.

In the second part of the analysis, we account for the fact that fawn survival outcomes are influenced by multiple factors, including regional variations. Specifically, we now model π, which determines the expected number of deaths as $m_i\pi$ through the relationship in (11), as a (generalized) linear function of the nominal covariate state_i, indicating the location where the fawns were observed (Colorado, Idaho, or Montana). We model state_i on the mean only, as it primarily captures location-driven shifts in baseline mortality (e.g., due to habitat, predator mix, or forage quality), whereas the dispersion (σ) and contamination (δ and η) parameters describe unobserved heterogeneity and tail inflation—reflecting an excess of extreme observations attributable to common factors such as study design and winter severity. Hence, these are treated as global noise features shared across states. The cBB-RM is specified as:

$$\begin{aligned}
\text{deaths}_i|\text{state}_i &\sim c\mathcal{BB}_{m_i}(\pi(\text{state}_i;\boldsymbol{\beta}), \sigma, \delta, \eta) \\
\text{logit}(\pi(\text{state}_i;\boldsymbol{\beta})) &= \beta_0 + \beta_{\text{Idaho}} I(\text{state}_i = \text{Idaho}) \\
&\quad + \beta_{\text{Montana}} I(\text{state}_i = \text{Montana}) \\
\log(\sigma) &= \alpha_0 \\
\text{logit}(\delta) &= \gamma_0 \\
\log(\eta - 1) &= \lambda_0
\end{aligned}$$

where $I(A)$ represents an indicator (dummy) variable that takes the value 1 if condition A is met and 0 otherwise, with Colorado serving as the reference category for state. As shown in Table 4,

TABLE 3 | Ranking of fitted models to the mule deer mortality data according to the AIC, BIC, and HQIC.

Model	#par	Log-likelihood	AIC	Rank	BIC	Rank	HQIC	Rank
B-RM	1	−234.424	470.849	4	472.107	4	471.211	4
BB-RM	2	−90.125	184.249	3	186.765	2	184.974	3
cBB-RM	4	−86.470	180.940	1	185.973	1	182.390	1
B2B-RM	3	−88.832	183.663	2	187.438	3	184.750	2

TABLE 4 | Ranking of fitted models to the mule deer mortality data with state as a covariate according to the AIC, BIC, and HQIC.

Model	#par	Log-likelihood	AIC	Rank	BIC	Rank	HQIC	Rank
B-RM	3	−224.930	445.860	4	459.634	4	456.947	4
BB-RM	4	−88.103	184.205	3	189.238	3	185.654	3
cBB-RM	6	−81.546	175.092	1	182.640	1	177.265	2
B2B-RM	5	−84.790	179.581	2	185.871	2	176.668	1

the cBB-RM outperforms the other RMs according to the AIC and BIC.

The LR test confirms that the cBB-RM model provides a statistically significant improvement over the BB-RM model, with a p-value of 0.001. Additionally, an LR test can be performed to compare the (null) cBB-D model against the (alternative) cBB-RM model, as the former represents a special case of the latter when $\beta_1 = \beta_2 = 0$. The resulting p-value of 0.007 indicates that incorporating state membership to model the mean count within the cBB framework leads to a significant improvement in model fit.

Table 5 shows the estimated regression coefficients for the BB and Cbb-RMs, along with their SEs in round brackets. Since a logit link function is used, the coefficients describe the change in log-odds of a mule deer dying from winter malnutrition. In the BB-RM, the intercept $\hat{\beta}_0 = -1.186$ (SE = 0.267) represents the log-odds of death in Colorado, corresponding to an odds of $e^{-1.186} \approx 0.31$ or probability of $\frac{0.31}{1+0.31} \approx 0.24$. The coefficient for Idaho is $\hat{\beta}_1 = -1.010$ (SE = 0.634), indicating that mule deer in Idaho have lower log-odds of death compared to Colorado, with an odds ratio of $e^{-1.010} \approx 0.36$, meaning about 64% lower odds of death. However, the SE is large, suggesting some uncertainty. The coefficient for Montana is $\hat{\beta}_2 = -0.774$ (SE = 0.560), implying an odds ratio of $e^{-0.774} \approx 0.46$, or about 54% lower odds of death compared to Colorado, though with a notable SE. In the cBB-RM, the intercept $\hat{\beta}_0 = -0.968$ (SE = 0.114) is slightly higher in absolute value, and the smaller SE suggests greater precision. The Idaho coefficient $\hat{\beta}_1 = -1.603$ (SE = 0.555) indicates a stronger reduction in log-odds compared to the BB-RM, with an odds ratio of $e^{-1.603} \approx 0.20$, meaning Idaho mule deer are about 80% less likely to die than those in Colorado. The Montana coefficient $\beta_2 = -0.281$ (SE = 0.333) is much smaller than in the BB model, corresponding to an odds ratio of $e^{-0.281} \approx 0.76$, meaning only 24% lower odds of death. The SE for Montana is also smaller, suggesting a more stable estimate. Overall, the cBB-RM suggests a stronger effect for Idaho and a weaker effect for Montana compared to the BB model, while also providing more precise estimates, likely due to better handling of extreme values and data heterogeneity. Specifically, since $\hat{\delta} = \text{logit}^{-1}(\hat{\gamma}_0) = 0.567$, approximately 56.7% of observations belong to the contaminant BB component, indicating an excess of extreme counts relative to the reference BB-RM. This percentage is lower than the 60.3% obtained in the initial analysis without covariates, suggesting that incorporating the state variable to explain the mean count reduces some of the excess variability previously attributed to extreme observations. In other words, state membership accounts for part of the overdispersion observed when no covariates were included. Furthermore, the estimated degree of contamination parameter is given by $\hat{\eta} = e^{\hat{\lambda}_0} + 1 = e^{4.606} + 1 \approx 100.8$, indicating that counts from the contaminant BB component are about 100 times more dispersed than those from the reference BB component. This suggests that while state membership helps explain some of the extreme observations, a substantial degree of contamination remains, emphasizing the necessity of using a flexible model like cBB-RM to properly capture the data heterogeneity.

TABLE 5 | Estimated coefficients and corresponding SEs (in brackets) of BB and cBB regression models to mule deer mortality data.

	Estimates (SEs)			
Parameter	**BB**		**cBB**	
β_0	−1.186	(0.267)	−0.968	(0.114)
β_{Idaho}	−1.010	(0.634)	−1.603	(0.555)
β_{Montana}	−0.774	(0.560)	−0.281	(0.333)
α_0	−1.427	(0.325)	−5.281	(1.373)
γ_0			0.269	(0.578)
λ_0			4.606	(1.457)

6.2 | Eurobarometer 95.1 Survey

Since the early 1970s, the European Commission has regularly surveyed public opinion in the Member States of the EU using the "Standard and Special Eurobarometer". The Eurobarometer

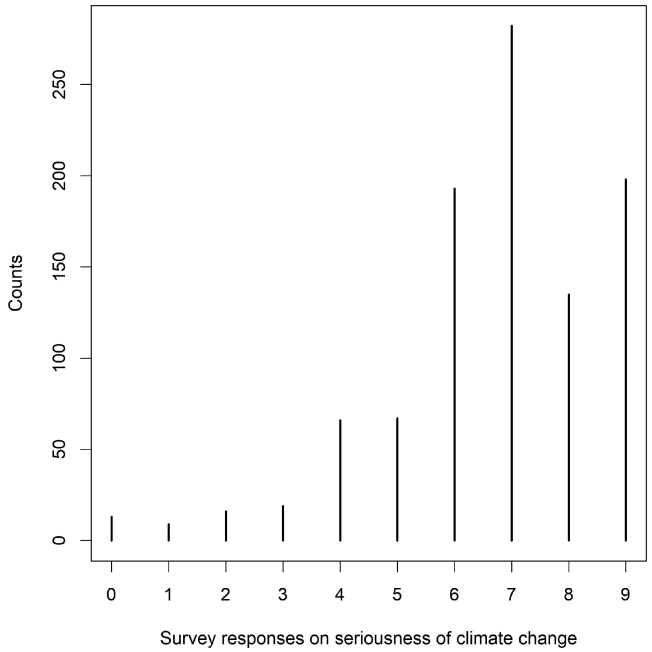


FIGURE 7 | Barplot of the counts of the survey responses on seriousness of climate change (0 = Not at all serious, 9 = Extremely serious).

95.1 survey was conducted in March–April 2021. For this analysis, we focus on responses from $n = 998$ survey participants in the Netherlands. The dataset is freely accessible at Commission, European and Brussels European Parliament (2023). Our response variable Y_i is the response of the seriousness of climate change out of 10, which we shifted by 1 so that '0' is instead considered as "not at all a serious problem" and "9," which is the m of the model, as "an extremely serious problem" (Sciandra et al. 2024). Hereafter, we shall also refer to Y as severity. A bar plot displaying the absolute frequency distribution of the observed Y-values is shown in Figure 7. Given the availability of demographic and socioeconomic data, we also analyze responses in relation to specific factors. In M_1, we begin by fitting the data without any covariates to investigate the suitability of the cBB-D. This initial approach assesses the responses on the perceived severity of climate change, providing a quantitative measure of individual concerns. In M_2, we include an explanatory variable on π, denoted as "politics", representing the political ideology of the respondents, measured by a scale where respondents rated themselves from 1 ("left") to 10 ("right"), while keeping the dispersion, proportion of extreme points and degree of contamination constant. M_3 is extended by including "age" as an explanatory variable on the overdispersion, while M_4 includes the settlement size where the respondent lives (either rural area, small town, or large town) as an explanatory variable of the degree of contamination. In formulas, the models are specified below.

M_1:

$$\begin{aligned}\text{severity}_i &\sim c\mathcal{BB}_9(\pi, \sigma, \delta, \eta)\\ \text{logit}(\pi) &= \beta_0\\ \log(\sigma) &= \alpha_0\\ \text{logit}(\delta) &= \gamma_0\\ \log(\eta - 1) &= \lambda_0\end{aligned}$$

M_2:

$$\begin{aligned}\text{severity}_i|\text{politics}_i &\sim c\mathcal{BB}_9(\pi(\text{politics}_i;\boldsymbol{\beta}), \sigma, \delta, \eta)\\ \text{logit}(\pi(\text{politics}_i;\boldsymbol{\beta})) &= \beta_0 + \beta_1\text{politics}_i\\ \log(\sigma) &= \alpha_0\\ \text{logit}(\delta) &= \gamma_0\\ \log(\eta - 1) &= \lambda_0\end{aligned}$$

M_3:

$$\begin{aligned}\text{severity}_i|\text{politics}_i, \text{age}_i &\sim c\mathcal{BB}_9(\pi(\text{politics}_i;\boldsymbol{\beta}), \sigma(\text{age}_i;\boldsymbol{\alpha}), \delta, \eta)\\ \text{logit}(\pi(\text{politics}_i;\boldsymbol{\beta})) &= \beta_0 + \beta_1\text{politics}_i\\ \log(\sigma(\text{age}_i;\boldsymbol{\alpha})) &= \alpha_0 + \alpha_1\text{age}_i\\ \text{logit}(\delta) &= \gamma_0\\ \log(\eta - 1) &= \lambda_0\end{aligned}$$

M_4:

$$\begin{aligned}\text{severity}_i|\text{politics}_i, \text{age}_i, \text{settlement}_i &\sim c\mathcal{BB}_9(\pi(\text{politics}_i;\boldsymbol{\beta}), \sigma(\text{age}_i;\boldsymbol{\alpha}),\\ &\qquad \delta, \eta(\text{settlement}_i;\lambda))\\ \text{logit}(\pi(\text{politics}_i;\boldsymbol{\beta})) &= \beta_0 + \beta_1\text{politics}_i\\ \log(\sigma(\text{age}_i;\boldsymbol{\alpha})) &= \alpha_0 + \alpha_1\text{age}_i\\ \text{logit}(\delta) &= \gamma_0\\ \log(\eta(\text{settlement}_i;\lambda) - 1) &= \lambda_0 + \lambda_1 I(\text{settlement}_i = \text{rural})\\ &\qquad + \lambda_2 I(\text{settlement}_i = \text{small town})\end{aligned}$$

for $i = 1, \ldots, 998$, with large town serving as the reference category for settlement size.

As shown in Table 6, M_4 performed best according to the AIC, whereas M_3 did so based on the BIC and HQIC. This is supported by the pairwise LR test p-values in Table 7, which indicate that M_4 does not represent a statistically significant improvement over M_3. Nevertheless, the parameter estimates and SEs for both M_3 and M_4 are reported in Table 8. The observed proportions and predicted probabilities for models M_1, M_2, M_3, and M_4 are displayed in Figure 8. Notably, M_2, M_3, and M_4 exhibit very similar behavior—consistent with the information criteria in Table 6—and provide more accurate predictions than M_1, as evidenced by their deviations from the observed proportions being smaller.

Due to the different parametrizations of the BB-RM, cBB-RM, and B2B model, direct comparisons for M_3 and M_4 are not possible, as certain parameters (e.g., the contamination degree η) are absent in the BB-RM and B2B model. However, comparisons can be made for M_1 and M_2, as seen in Tables 9 and 10, respectively.

From Table 9, we observe that the cBB-RM performs best according to the AIC and HQIC, whereas the B2B model performs better according to the BIC. However, this is not the case for M_2, where

TABLE 6 | Ranking of fitted cBB-RMs to the Netherlands Survey Responses on Climate Change data according to the AIC, BIC, and HQIC.

Model	#par	Log-likelihood	AIC	Rank	BIC	Rank	HQIC	Rank
M_1	4	−1890.949	3789.899	4	3809.522	4	3797.357	4
M_2	5	−1809.903	3629.807	3	3654.336	3	3639.130	3
M_3	6	−1805.580	3623.159	2	3652.594	1	3634.347	1
M_4	8	−1803.433	3622.866	1	3662.112	2	3637.783	2

TABLE 7 | Pairwise likelihood ratio test p-values of fitted cBB-RMs to the Netherlands Survey Responses on Climate Change data.

	M_1	M_2	M_3	M_4
M_1				
M_2	0.000			
M_3	0.000	0.003		
M_4	0.000	0.004	0.117	

TABLE 8 | Maximum likelihood estimates and corresponding standard errors (in brackets) for models M_3 and M_4 fitted to the Netherlands Survey Responses on Climate Change data.

Parameter	M_3		M_4	
β_0	2.146	(0.095)	2.131	(0.094)
β_1	−0.213	(0.016)	−0.210	(0.016)
α_0	−6.997	(1.376)	−7.740	(1.483)
α_1	0.045	(0.017)	0.057	(0.021)
γ_0	−1.450	(0.317)	−1.499	(0.274)
λ_0	4.840	(0.820)	6.131	(1.130)
λ_1			−1.435	(0.934)
λ_2			−1.343	(0.850)

the B2B model outperformed the cBB-RM according to all the criteria, as seen in Table 10. Despite this, the cBB-RM retains the advantage of having interpretable parameters. Moreover, Figure 9 indicates that the cBB and B2B RMs exhibit comparable performance and yield similar predictions.

Table 11 presents the estimated regression coefficients under M_2 for the BB-RM and cBB-RM, along with their standard errors in parentheses. Since a logit link function is used for the parameter π, the coefficients describe the change in log-odds of the outcome variable as a function of political ideology. The estimated coefficient for political ideology is $\hat{\beta}_1 = -0.235$ (SE = 0.017), indicating that as respondents move one unit to the right on the political scale, their log-odds of the outcome decrease. This corresponds to an odds ratio of $e^{-0.235} \approx 0.79$, meaning that for each one-point increase in political ideology (toward a more right-leaning stance), the odds of the outcome decrease by approximately 21%. The relatively small SE suggests a precise estimate. In the cBB-RM, the intercept $\hat{\beta}_0 = 2.132$ (SE = 0.101) is slightly lower, suggesting a similar but slightly reduced baseline log-odds. The political ideology coefficient $\hat{\beta}_1 = -0.210$ (SE = 0.018) is also slightly smaller in magnitude compared to the BB-RM, with an odds ratio of $e^{-0.210} \approx 0.81$, meaning a slightly weaker association between political ideology and the outcome. The SE remains small. The dispersion parameter α_0 in the BB-RM is estimated at −2.239 (SE = 0.105), whereas in the cBB model, it is −5.406 (SE = 4.560), indicating a large shift with greater uncertainty in the contaminated model. The estimated proportion of observations belonging to the contaminant BB component is given by $\hat{\delta} = \text{logit}^{-1}(\hat{\gamma}_0) = \text{logit}^{-1}(-1.271) \approx 0.22$, meaning about 22% of excess of extreme responses. The degree of contamination is estimated as $\hat{\eta} = e^{\hat{\lambda}_0} + 1 = e^{5.430} + 1 \approx 229.15$, indicating that the contaminant BB component exhibits a dispersion more than 200 times greater than that of the reference BB component. This suggests that while political ideology explains some of the variability in responses, a significant amount of overdispersion remains, reinforcing the importance of using a flexible model like cBB-RM to accommodate extreme values.

As noted in Table 8, under M_3, the estimated mean parameters remain close to those under M_2, indicating a consistent relationship between political ideology and the outcome. However, the SE of the mean coefficients decreases compared to those under M_2. The dispersion parameter now varies with age: $\hat{\alpha}_0 = -6.997$ (SE = 1.376) and $\hat{\alpha}_1 = 0.045$ (SE = 0.017). Since dispersion is modelled on the log scale, this suggests that for each additional year of age, dispersion increases by a multiplicative factor of $e^{0.045} \approx 1.05$, meaning older respondents exhibit slightly higher overdispersion. The SE for age is small, indicating a stable estimate. In M_4, settlement size is introduced as a predictor of the degree of contamination while keeping the dispersion-age relationship from M_3. The estimated mean parameters remain similar, and the dispersion parameters are also close to those in M_3, with $\hat{\alpha}_0 = -7.740$ (SE = 1.483) and $\hat{\alpha}_1 = 0.057$ (SE = 0.021), indicating a slightly stronger effect of age on dispersion. However, settlement size has a notable effect on the degree of contamination. The baseline degree of contamination is estimated as $\hat{\lambda}_0 = 6.131$ (SE = 1.130), corresponding to $\hat{\eta} = e^{6.131} + 1 \approx 460.2$, indicating substantial contamination in the reference category (large towns). For rural respondents, $\hat{\lambda}_1 = -1.435$ (SE = 0.934), leading to a degree of contamination of $\hat{\eta}_{\text{rural}} = e^{6.131-1.435} + 1 \approx 139.7$, while for small-town respondents, $\hat{\lambda}_2 = -1.343$ (SE = 0.850), giving $\hat{\eta}_{\text{small town}} = e^{6.131-1.343} + 1 \approx 158.5$. This suggests that contamination is substantially lower in rural and small-town areas compared to large towns.

Overall, these findings highlight the benefits of the cBB framework in handling overdispersion and an excess of extreme observations. The inclusion of covariates for dispersion and contamination enhances model flexibility and provides deeper insights into the factors influencing the response variability.

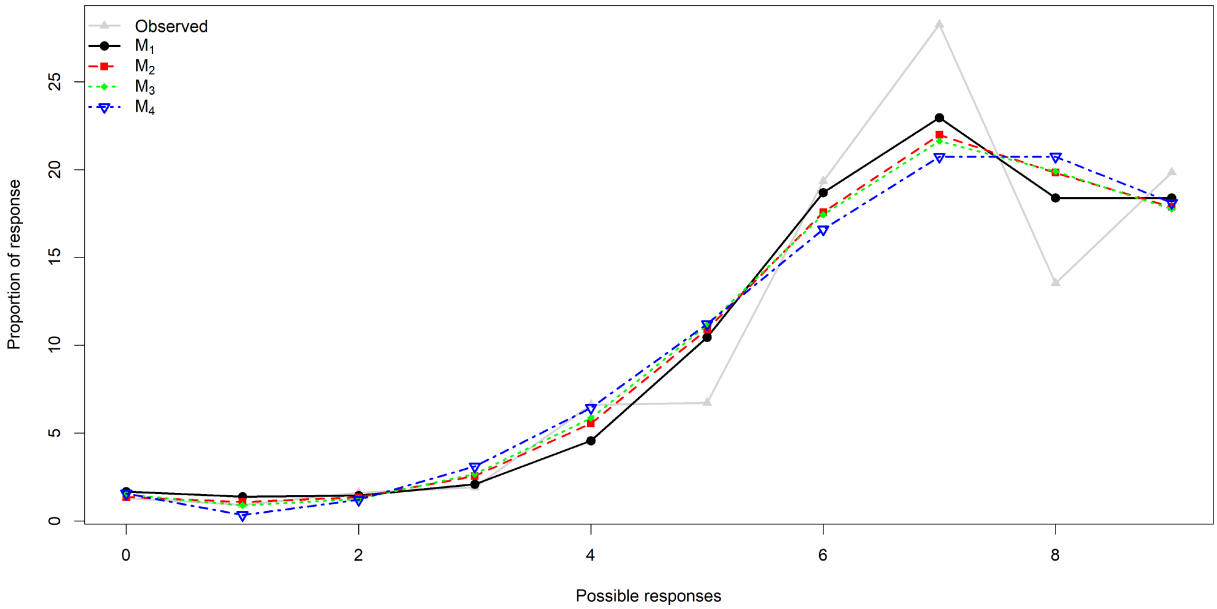


FIGURE 8 | Observed proportions and predicted probabilities for the fitted cBB-RM's on the survey responses regarding the perceived seriousness of climate change, with covariates specified as in M1, M2, M3, and M4.

TABLE 9 | Ranking of fitted models to the Netherlands Survey Responses on Climate Change data according to the AIC, BIC, and HQIC, using no covariates as specified in M_1.

Model	#par	Log-likelihood	AIC	Rank	BIC	Rank	HQIC	Rank
B-D	1	−2133.422	4268.843	4	4273.749	4	4270.708	4
BB-D	2	−1937.026	3878.051	3	3887.863	3	3881.781	3
cBB-D	4	−1890.949	3789.899	1	3809.522	2	3797.357	1
B2B-D	3	−1893.708	3793.416	2	3808.133	1	3799.010	2

TABLE 10 | Ranking of fitted models to the Netherlands Survey Responses on Climate Change data according to the AIC, BIC, and HQIC, using covariates as specified in M_2.

Model	#par	Log-likelihood	AIC	Rank	BIC	Rank	HQIC	Rank
B-RM	2	−1951.782	3907.564	4	3917.375	4	3911.293	4
BB-RM	3	−1843.396	3692.792	3	3707.509	3	3698.386	3
cBB-RM	5	−1809.903	3629.807	2	3654.336	2	3639.130	2
B2B-RM	4	−1809.377	3626.755	1	3646.378	1	3630.349	1

7 | Conclusion

In this paper, we introduce the cBB-D as a new approach for modeling bounded count data. Our model is formulated as a simple mixture of two beta-binomial distributions (BB-Ds) that share the same mean but differ in dispersion; advantageously, this implies a closed-form expression of the PMF. The primary motivation behind this formulation is to protect the reference BB-D—characterized by lower dispersion—from model misspecification, particularly due to an excess of extreme observations. These "additional" extreme observations are assumed to arise from the contaminant BB-D, which exhibits higher dispersion in our mixture framework. We place particular emphasis on the flexibility in capturing key characteristics of interest, including mean, variance, skewness, and kurtosis. Another notable advantage of the cBB-D is that its formulation introduces only two additional (contamination) parameters with respect to the reference BB-D, both of which have intuitive and practical interpretations. These parameters represent the proportion of observations from the contaminant BB-D and the degree of contamination, where the latter quantifies the extent to which the contaminant BB-D is more dispersed than the reference BB-D.

Another strength of the cBB-D is its parameterization, which includes a mean-related parameter and a dispersion parameter in addition to the contamination parameters. This structure allows the cBB-D to be seamlessly integrated into a regression framework, giving rise to the cBB-RM. Notably, this regression

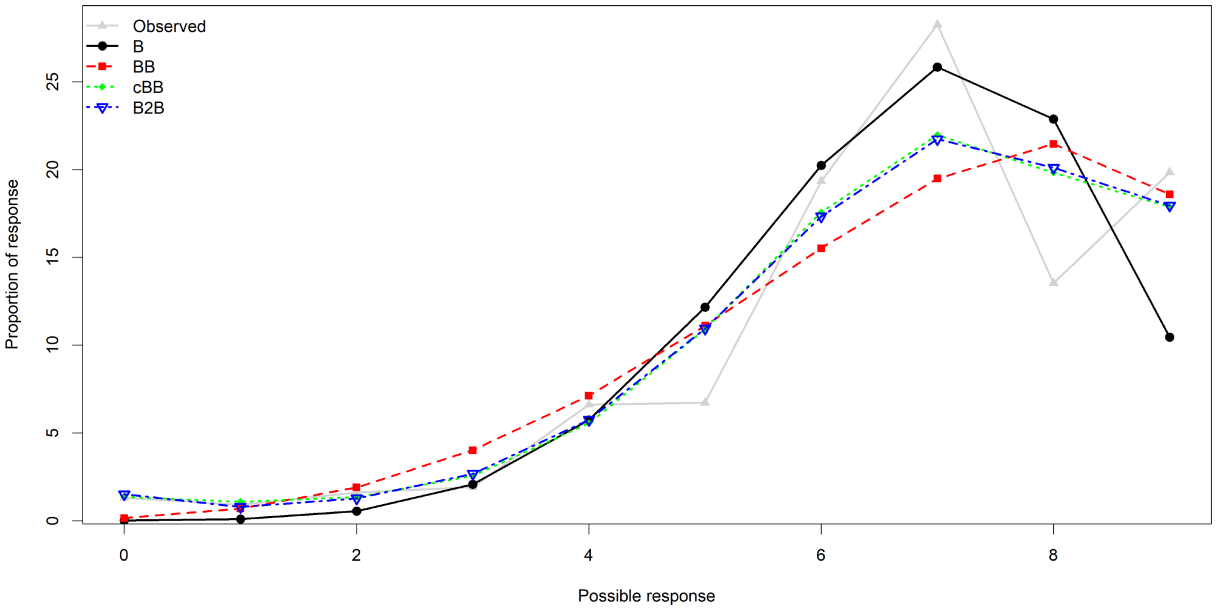


FIGURE 9 | Observed proportions and predicted probabilities for the fitted regression models on the survey responses regarding the perceived seriousness of climate change, with covariates specified as in M_2.

TABLE 11 | Maximum likelihood estimates and corresponding standard errors (in brackets) for model M_2 fitted to the Netherlands Survey Responses on Climate Change data.

Parameter	BB		cBB	
β_0	2.281	(0.096)	2.132	(0.101)
β_1	−0.235	(0.017)	−0.210	(0.018)
α_0	−2.239	(0.105)	−5.406	(4.560)
γ_0			−1.271	(0.512)
λ_0			5.430	(4.168)

framework is not restricted to modeling the mean only; rather, it extends to all parameters of the cBB-D, enabling the inclusion of different covariates for each parameter. This results in a highly flexible RM for bounded count response variables, particularly in the presence of an excess of extreme values.

From an inferential standpoint, we propose an EM algorithm for the ML estimation of the cBB-RM parameters. The robustness of the model to outliers—referred to in this paper as extreme values—is examined through a sensitivity analysis, assessing their potential to introduce bias in the estimated regression parameters. The real-world applications—ranging from modeling species counts influenced by environmental factors to analyzing public opinion on climate-related issues—further demonstrate the effectiveness of our approach, highlighting its advantages over established RMs for bounded count data. These results position the cBB-RM as a compelling alternative for analyzing bounded count data characterized by an excess of extreme observations.

Future work could focus on implementing the cBB-RM within the generalized additive model for location, scale, and shape (GAMLSS) framework (Stasinopoulos et al. 2017), following the approach of Bayes et al. (2024). Moreover, the impact of alternative link functions on model performance could be investigated, leveraging the flexibility of GAMLSS to specify distinct link functions for each distributional parameter.

Acknowledgments

Arno Otto acknowledges the support from the DSI-NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS) under grant no. 2025-030-STA-Hidden Patterns, as well as by the UCDP and DRI of the University of Pretoria. Johan Ferreira and Andriette Bekker have been partially supported by the National Research Foundation (NRF) of South Africa (SA), grant RA201125576565, Nr. 145681; RA171022270376, grant Nr. 119109; and grant SRUG2204203865 No. 120839. The opinions expressed and conclusions arrived at are those of the authors and are not necessarily to be attributed to the NRF. Antonio Punzo acknowledges the support by the Italian Ministry of University and Research (MUR) under the PRIN 2022 grant number 2022XRHT8R (CUP: E53D23005950006), as part of 'The SMILE Project: Statistical Modelling and Inference to Live the Environment', funded by the European Union – Next Generation EU. Cristina Tortora has been partially supported by NSF grant No. 2209974.

Funding

This work was supported by the DSI-NRF Centre of Excellence in Mathematical and Statistical Sciences (Grant No. 2025-030-STA-Hidden Patterns), the National Research Foundation of South Africa (Grant Nos. RA201125576565 Nr. 145681, RA171022270376 Nr. 119109, RUG2204203865 nr. 120839), the Italian Ministry of University and Research (MUR) (PRIN 2022 Grant No. 2022XRHT8R (CUP:22E53D23005950006)), and the National Science Foundation (Grant Nr. 2209974).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All datasets considered in this paper are freely available on the internet.

References

Akaike, H. 1974. "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control* 19, no. 6: 716–723.

Arikan, G., and D. Günay. 2021. "Public Attitudes Towards Climate Change: A Cross-Country Analysis." *British Journal of Politics and International Relations* 23, no. 1: 158–174.

Arostegui, I., A. Padierna, and J. M. Quintana. 2010. "Assessment of HRQoL in Patients With Eating Disorders by the Beta-Binomial Regression Approach." *International Journal of Eating Disorders* 43, no. 5: 455–463.

Aspinall, R., and K. Matthews. 1994. "Climate Change Impact on Distribution and Abundance of Wildlife Species: An Analytical Approach Using GIS." *Environmental Pollution* 86, no. 2: 217–223.

Bayes, C. L. S. A. I. G., J. L. Bazán, and L. Valdivieso. 2024. "A Robust Regression Model for Bounded Count Health Data." *Statistical Methods in Medical Research* 33, no. 8: 1392–1411.

Commission, European and Brussels European Parliament. 2023. "Eurobarometer 95.1 (2021). GESIS, Cologne. ZA7781 Data file Version 2.0.0," https://doi.org/10.4232/1.14079.

Davies, L., and U. Gather. 1993. "The Identification of Multiple Outliers." *Journal of the American Statistical Association* 88, no. 423: 782–792.

Gallop, R. J., R. H. Rieger, S. McClintock, and D. C. Atkins. 2013. "A Model for Extreme Stacking of Data at Endpoints of a Distribution: Illustration With W-Shaped Data." *Statistical Methodology* 10, no. 1: 29–45.

Griffiths, D. A. 1973. "Maximum Likelihood Estimation for the Beta-Binomial Distribution and an Application to the Household Distribution of the Total Number of Cases of a Disease." *Biometrics* 29: 637–648.

Hannan, E. J., and B. G. Quinn. 1979. "The Determination of the Order of an Autoregression." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 41, no. 2: 190–195.

Keller-Ressel, M. 2022. "W-Shaped Implied Volatility Curves in a Variance-Gamma Mixture Model," arXiv preprint arXiv:2209.14726.

Konisky, D. M., and N. D. Woods. 2010. "Exporting Air Pollution? Regulatory Enforcement and Environmental Free Riding in the United States." *Political Research Quarterly* 63, no. 4: 771–782.

Korkmaz, M. Ç. 2020. "A New Heavy-Tailed Distribution Defined on the Bounded Interval: The Logit Slash Distribution and Its Application." *Journal of Applied Statistics* 47, no. 12: 2097–2119.

Ley, C., S. Babić, and D. Craens. 2021. "Flexible Models for Complex Data With Applications." *Annual Review of Statistics and Its Application* 8, no. 1: 369–391.

Martin, B. D., D. Witten, and A. D. Willis. 2020. "Modeling Microbial Abundances and Dysbiosis With Beta-Binomial Regression." *Annals of Applied Statistics* 14, no. 1: 94–115.

Muluneh, M. G. 2021. "Impact of Climate Change on Biodiversity and Food Security: A Global Perspective a Review Article." *Agriculture & Food Security* 10, no. 1: 1–25.

Otto, A. F., J. T. Ferreira, A. Bekker, A. Punzo, and S. D. Tomarchio. 2025. "A Refreshing Take on the Inverted Dirichlet via a Mode Parameterization With Some Statistical Illustrations." *Journal of the Korean Statistical Society* 54, no. 1: 314–341.

Otto, A. F., J. T. Ferreira, S. D. Tomarchio, A. Bekker, and A. Punzo. 2025. "A Contaminated Regression Model for Count Health Data." *Statistical Methods in Medical Research* 34, no. 2: 369–389.

Paul, S., and K. K. Saha. 2007. "The Generalized Linear Model and Extensions: A Review and Some Biological and Environmental Applications." *Environmetrics* 18, no. 4: 421–443.

Punzo, A. 2019. "A New Look at the Inverse Gaussian Distribution With Applications to Insurance and Economic Data." *Journal of Applied Statistics* 46, no. 7: 1260–1287.

Punzo, A., and L. Bagnato. 2021. "Modeling the Cryptocurrency Return Distribution via Laplace Scale Mixtures." *Physica A* 563, no. 1: 125354.

Punzo, A., and L. Bagnato. 2025. "Asymmetric Laplace Scale Mixtures for the Distribution of Cryptocurrency Returns." *Advances in Data Analysis and Classification* 19, no. 4: 275–322.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. http://www.R-project.org/.

Ryan, D. A. J. 2007. "Application of the Beta-Binomial Model for the Detection of Rare Marine Benthos Using Point Intercept Techniques." *Environmetrics* 18, no. 4: 361–373.

Schwarz, G. 1978. "Estimating the Dimension of a Model." *Annals of Statistics* 6, no. 2: 461–464.

Sciandra, M., S. Fasola, A. Albano, C. Di Maria, and A. Plaia. 2024. "Discrete Beta and Shifted Beta-Binomial Models for Rating and Ranking Data." *Environmental and Ecological Statistics* 31, no. 2: 317–338.

Spina, F. 2015. "Environmental Justice and Patterns of State Inspections." *Social Science Quarterly* 96, no. 2: 417–429.

Stasinopoulos, M. D., R. A. Rigby, G. Z. Heller, V. Voudouris, and F. De Bastiani. 2017. *Flexible Regression and Smoothing: Using GAMLSS in R*. CRC Press, Taylor & Francis Group.

Tomarchio, S. D., A. Punzo, J. T. Ferreira, and A. Bekker. 2025. "A New Look at the Dirichlet Distribution: Robustness, Clustering, and Both Together." *Journal of Classification* 42: 31–53.

Unsworth, J. W., D. F. Pac, G. C. White, and R. M. Bartmann. 1999. "Mule Deer Survival in Colorado, Idaho, and Montana." *Journal of Wildlife Management* 63, no. 1: 315–326.

Wagener, M., A. Bekker, M. Arashi, and A. Punzo. 2024. "Uncovering a Generalised Gamma Distribution: From Shape to Interpretation." *Results in Applied Mathematics* 22: 100461.

Yee, S. H., D. L. Santavy, and M. G. Barron. 2008. "Comparing Environmental Influences on Coral Bleaching Across and Within Species Using Clustered Binomial Regression." *Ecological Modelling* 218, no. 1–2: 162–174.

Appendix A

Proofs

Proof. Proof Proposition 1 The cBB distribution in (10) has the hierarchical representation

$$\begin{aligned} W &\sim \mathcal{TP}_{\{1,\eta\}}(\delta) \\ Y|W = w &\sim \mathcal{BB}_m(\pi, w\sigma), \end{aligned} \tag{A1}$$

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

$$W = \begin{cases} 1 & \text{with probability } 1-\delta, \\ \eta & \text{with probability } \delta. \end{cases} \tag{A2}$$

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

a. if $\delta \to 0^+$, from (A2) it follows that $W \xrightarrow{D} 1$ and, therefore, according to (A1–A2), $Y \xrightarrow{D} \mathcal{BB}_m(\pi, \sigma)$;

b. if $\eta \to 1^+$, from (A2) it follows that $W \xrightarrow{D} 1$ and, as before, according to (A1–A2), $Y \xrightarrow{D} \mathcal{BB}_m(\pi, \sigma)$;

c. if $\delta \to 0^+$ and $\sigma \to 0^+$, from the proof for (a) and from the results given in Griffiths (1973), it follows that $Y \xrightarrow{D} \mathcal{B}_m(\pi)$;

d. if $\eta \to 1^+$ and $\sigma \to 0^+$, from the proof for (b) and, again, as demonstrated in Griffiths (1973), it follows that $Y \xrightarrow{D} \mathcal{B}_m(\pi)$. □

Proposition 2. *Characteristics of the cBB distribution. If* $Y \sim c\mathcal{BB}_m(\pi, \sigma, \delta, \eta)$, *then*

a. the variance is given by

$$\mathrm{Var}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta) = \frac{m\pi(1-\pi)[(1-\delta)(1+m\sigma)(1+\eta\sigma)+\delta(1+m\eta\sigma)(1+\sigma)]}{(1+\sigma)(1+\eta\sigma)};$$

b. the skewness is given by

$$\mathrm{Skew}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta) = \frac{1-\delta}{\left(\mathrm{Var}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta)\right)^{\frac{3}{2}}}\left(\frac{m\pi(1-\pi)(1-2\pi)(1+m\sigma)(1+2m\sigma)}{(1+\sigma)(1+2\sigma)}\right) + \frac{\delta}{\left(\mathrm{Var}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta)\right)^{\frac{3}{2}}}\left(\frac{m\pi(1-\pi)(1-2\pi)(1+m\sigma\eta)(1+2m\sigma\eta)}{(1+\sigma\eta)(1+2\sigma\eta)}\right);$$

c. the kurtosis is given by

$$\mathrm{ExKurt}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta) = -3 + \frac{m\pi(1-\pi)}{\left[\mathrm{Var}_{\mathrm{cBB}}(Y;\pi,\sigma,\delta,\eta)\right]^2} \times \left(\frac{(1-\delta)(m\sigma+1)(6(3(\pi-1)\pi+1)m^2\sigma^2+3m\sigma(2-(\pi-1)\pi(m-6))-3(\pi-1)\pi(m-2)-\sigma+1)}{(\sigma+1)(2\sigma+1)(3\sigma+1)} + \frac{\delta(\eta m\sigma+1)(-\eta\sigma+6\eta^2(3(\pi-1)\pi+1)m^2\sigma^2+3\eta m\sigma(2-(\pi-1)\pi(m-6))-3(\pi-1)\pi(m-2)+1)}{(\eta\sigma+1)(2\eta\sigma+1)(3\eta\sigma+1)}\right)$$

Proof. If $Y \sim c\mathcal{BB}_m(\pi, \sigma, \delta, \eta)$, then from the hierarchical representation in (A1–A2),

a.

$$\begin{aligned}\mathrm{Var}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta) &= \mathrm{E}_{\mathrm{cBB}_m}\left(Y^2;\pi,\sigma,\delta,\eta\right) - \left[\mathrm{E}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta)\right]^2 \\ &= \mathrm{E}_{\mathrm{TP}}\left[\mathrm{Var}_{\mathrm{BB}_m}(Y|W=w;\pi,\sigma);\delta\right] \\ &\quad + \mathrm{Var}_{\mathrm{TP}}\left[\mathrm{E}_{\mathrm{BB}_m}(Y|W=w;\pi,\sigma);\delta\right]\end{aligned}$$

where

$$\mathrm{Var}_{\mathrm{BB}_m}(Y|W=w;\pi,\sigma) = \frac{m\pi(1-\pi)(1+mw\sigma)}{1+w\sigma}.$$

Since $\mathrm{E}_{\mathrm{BB}_m}(Y|W=w;\pi,\sigma) = m\pi$, it follows that

$$\begin{aligned}&\mathrm{Var}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta) \\ &= \mathrm{E}_{\mathrm{TP}}\left[\frac{m\pi(1-\pi)(1+mw\sigma)}{1+w\sigma};\delta\right] \\ &= (1-\delta)\frac{m\pi(1-\pi)(1+m\sigma)}{1+\sigma} + \delta\frac{m\pi(1-\pi)(1+m\eta\sigma)}{1+\eta\sigma} \\ &= \frac{m\pi(1-\pi)[(1-\delta)(1+m\sigma)(1+\eta\sigma)+\delta(1+m\eta\sigma)(1+\sigma)]}{(1+\sigma)(1+\eta\sigma)}.\end{aligned}$$

b.

$$\begin{aligned}&\mathrm{Skew}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta) \\ &= \mathrm{E}_{\mathrm{cBB}_m}\left[\left(\frac{Y-\mathrm{E}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta)}{\sqrt{\mathrm{Var}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta)}}\right)^3;\pi,\sigma,\delta,\eta\right] \\ &= \sum_{i=0}^{m}\left(\frac{y_i-\mathrm{E}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta)}{\sqrt{\mathrm{Var}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta)}}\right)^3 f_{\mathrm{cBB}_m}(y_i;\pi,\sigma,\delta,\eta) \\ &= \sum_{w}\sum_{i=0}^{m}\left(\frac{y_i-m\pi}{\sqrt{\mathrm{Var}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta)}}\right)^3 f_{\mathrm{BB}_m}(y_i;\pi,\sigma w) \\ &= \frac{1}{\left(\mathrm{Var}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta)\right)^{\frac{3}{2}}}\sum_{w}\sum_{i=0}^{m}(y_i-m\pi)^3 f_{\mathrm{BB}_m}(y_i;\pi,\sigma w) \\ &= \frac{1}{\left(\mathrm{Var}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta)\right)^{\frac{3}{2}}}\sum_{w}\left(\frac{m\pi(1-\pi)(2\pi-1)(1+m\sigma w)(1+2m\sigma w)}{(1+\sigma w)(1+2\sigma w)}\right) \\ &= \frac{1-\delta}{\left(\mathrm{Var}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta)\right)^{\frac{3}{2}}}\left(\frac{m\pi(1-\pi)(1-2\pi)(1+m\sigma)(1+2m\sigma)}{(1+\sigma)(1+2\sigma)}\right) \\ &\quad + \frac{\delta}{\left(\mathrm{Var}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta)\right)^{\frac{3}{2}}}\left(\frac{m\pi(1-\pi)(1-2\pi)(1+m\sigma\eta)(1+2m\sigma\eta)}{(1+\sigma\eta)(1+2\sigma\eta)}\right).\end{aligned}$$

c.

$$\begin{aligned}&\mathrm{Kurt}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta) \\ &= \mathrm{E}_{\mathrm{cBB}_m}\left[\left(\frac{Y-\mathrm{E}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta)}{\sqrt{\mathrm{Var}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta)}}\right)^4;\pi,\sigma,\delta,\eta\right] \\ &= \sum_{i=0}^{m}\left(\frac{y_i-\mathrm{E}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta)}{\sqrt{\mathrm{Var}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta)}}\right)^4 f_{\mathrm{cBB}_m}(y_i;\pi,\sigma,\delta,\eta) \\ &= \sum_{w}\sum_{i=0}^{m}\left(\frac{y_i-m\pi}{\sqrt{\mathrm{Var}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta)}}\right)^4 f_{\mathrm{BB}_m}(y_i;\pi,\sigma w) \\ &= \sum_{w}\frac{1}{\left[\mathrm{Var}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta)\right]^2}\sum_{i=0}^{m}(y_i-m\pi)^4 f_{\mathrm{BB}_m}(y_i;\pi,\sigma w)\end{aligned}$$

where

$$\sum_{i=0}^{m}\left(y_i - m\pi\right)^4 f_{\mathrm{BB}_m}(y_i;\pi,\sigma w) = \frac{m\pi(1-\pi)(m\sigma w+1)(6(3(\pi-1)\pi+1)m^2\sigma^2w^2 - 3(\pi-1)\pi(m-2)+3m\sigma w(2-(\pi-1)\pi(m-6))-\sigma w+1)}{(\sigma w+1)(2\sigma w+1)(3\sigma w+1)}.$$

Thus,

$$\begin{aligned}\mathrm{ExKurt}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta) = -3 + \frac{m\pi(1-\pi)}{\left[\mathrm{Var}_{\mathrm{cBB}_m}(Y;\pi,\sigma,\delta,\eta)\right]^2} \\ \times\left(\frac{(1-\delta)(m\sigma+1)(6(3(\pi-1)\pi+1)m^2\sigma^2+3m\sigma(2-(\pi-1)\pi(m-6))-3(\pi-1)\pi(m-2)-\sigma+1)}{(\sigma+1)(2\sigma+1)(3\sigma+1)}\right. \\ \left. + \frac{\delta(\eta m\sigma+1)(-\eta\sigma+6\eta^2(3(\pi-1)\pi+1)m^2\sigma^2+3\eta m\sigma(2-(\pi-1)\pi(m-6))-3(\pi-1)\pi(m-2)+1)}{(\eta\sigma+1)(2\eta\sigma+1)(3\eta\sigma+1)}\right).\end{aligned}$$

□

Appendix B

A Practitioner's Guide to the cBB Package

This appendix serves as a "practitioner's guide" to implementing the methodology presented in this paper. It provides details on the **cBB** package for R, along with examples that reproduce the results in Section 6.1, and illustrate its broader application. The package is available on GitHub at https://github.com/arnootto/cBB.

Installation

To install the **cBB** package directly from GitHub, use the following code in R:

```
#install.packages("devtools")
library(devtools)
install_github("arnootto/cBB")
```

Main function: `ml.cmbb()`

The `ml.cmbb()` function performs ML estimation of the cBB-RM. Below is a detailed description of its arguments and return values:

```
ml.cmbb(
  formula,
  sigma.formula = ~1,
  delta.formula = ~1,
  eta.formula = ~1,
  data,
  init = NULL,
  method = "BFGS",
  reltol = 1e-15,
  maxit = 10000,
  hessian = TRUE,
  EM = TRUE
)
```

Arguments

formula	An object of class formula: a symbolic description of the probability of success parameter (pi).
sigma.formula	A formula for the dispersion parameter (sigma). Defaults to ~1.
delta.formula	A formula for the proportion of extreme values parameter (delta). Defaults to ~1.
eta.formula	A formula for the inflation parameter (eta). Defaults to ~1.
data	A mandatory data frame containing the variables in the model.
init	Optional vector of initial values. If NULL, values are initialized using ml.mbb with delta = 0.01 and eta = 1.1.
method	Optimization method to be used. Default is "BFGS"; "Nelder-Mead" is also supported.
reltol	Relative convergence tolerance in optimization. Default is 1e-15.
maxit	Maximum number of iterations for the optimizer. Default is 1000.
hessian	Logical; if TRUE, computes the Hessian matrix for standard errors. Default is TRUE.
EM	Logical; if TRUE, the EM algorithm is used for maximum likelihood estimation. If FALSE, direct numerical optimization is used. Default is TRUE.

Return Values

The ml.cmbb() function returns a list with the following components:

results	A data frame with parameter estimates, standard errors, *t*-values, and *p*-values (if hessian = TRUE).
beta	Maximum likelihood estimates of the regression coefficients for pi.
alpha	Maximum likelihood estimates of the regression coefficients for sigma.
gamma	Maximum likelihood estimates of the regression coefficients for delta.
lambda	Maximum likelihood estimates of the regression coefficients for eta.
pi	Fitted values of the probability of success parameter.
sigma	Fitted values of the dispersion parameter.
delta	Fitted values of the contamination proportion.
eta	Fitted values of the inflation parameter.
X	The design matrix for pi.
U	The design matrix for sigma.
V	The design matrix for delta.
Z	The design matrix for eta.
y	The response variable.
loglike	The log-likelihood value at convergence.
AIC	Akaike Information Criterion.
BIC	Bayesian Information Criterion.
HQIC	Hannan-Quinn Information Criterion.

Example

The following code reproduces the BB-RM and cBB-RM fit in the mule deer example in Section 6.1.

```
library(cBB)
data("MuleDeer")

#intercept only regression models
est_mbb <- ml.mbb(formula = cbind(Winter_malnutrition_n, Radiocollared_fawns -
  Winter_malnutrition_n) ~ 1, data = MuleDeer, reltol = 1e-10, method =
  "Nelder-Mead")
est_cmbb <- ml.cmbb(formula = cbind(Winter_malnutrition_n, Radiocollared_fawns -
  Winter_malnutrition_n) ~ 1, data = MuleDeer, reltol = 1e-10, method =
  "Nelder-Mead")
est_mbb$AIC #returns AIC value of fitted BB-RM
est_mbb$BIC #returns BIC value of fitted BB-RM
est_cmbb$AIC #returns AIC value of fitted cBB-RM
est_cmbb$BIC #returns BIC value of fitted cBB-RM

#state as covariate
est_mbb <- ml.mbb(formula = cbind(Winter_malnutrition_n, Radiocollared_fawns -
  Winter_malnutrition_n) ~ State, data = MuleDeer, reltol = 1e-10, method =
  "BFGS")
est_cmbb <- ml.cmbb(formula = cbind(Winter_malnutrition_n, Radiocollared_fawns -
  Winter_malnutrition_n) ~ State, data = MuleDeer, reltol = 1e-10, method =
  "BFGS")
est_mbb$AIC #returns AIC value of fitted BB-RM
est_mbb$BIC #returns BIC value of fitted BB-RM
est_cmbb$AIC #returns AIC value of fitted cBB-RM
est_cmbb$BIC #returns BIC value of fitted cBB-RM
```

Other functions in the cBB package

The **cBB** package also includes the following functions:

`dcbetabinom()`	Computes the PDF of the cBB-D.
`dmbetabinom()`	Computes the PDF of the BB-D.
`ml.mbb()`	ML estimation for the BB-RM.
`rcmbb_regression()`	Generates synthetic data from a cBB-RM.
`rmbb_regression()`	Generates synthetic data from a BB-RM.

For more details, use the `help` command or by typing `?function_name` in R (e.g., ?dcbetabinom).