# Estimation of contact matrices with a new count data model and surveys

Solym Mawaki Manou-Abi[1*†], Essoham Ali[2†], Yousri Slaoui[1], Julien Balicchi[3,4]

[2]Laboratoire , University of ..., XXX, XXXX, France.
[1]Laboratoire de Mathématiques et Applications, Université de Poitiers, Poitiers, UMR CNRS 7348, France.

*Corresponding author(s). E-mail(s): solym.Manou.abi@math.univ-poitiers.fr;
Contributing authors: essoham.ali@univ-ubs.fr; .....; .....;
†These authors contributed equally to this work.

**Abstract**

This article presents a statistical analysis of contact rate matrices for the island of Mayotte, using contact surveys and local demographic data. The objective is to estimate these matrices while accounting for specific data characteristics, such as excess zeros and outliers, which are modeled through the Zero-Inflated Bell (ZIBell) model. This model combines a component for structural zeros and a Bell-type distribution for other observations. Parameter estimation is performed via the Minimum Density Power Divergence Estimator (MDPDE), which is robust to contaminated data while ensuring the consistency and asymptotic normality of the estimator. Simulation results demonstrate the performance and robustness of this estimator in the presence of contamination, as compared to the maximum likelihood estimator.

## 1 Introduction

### 1.1 Statistical models for count data

Analyzing count data has become a significant focus in statistical modeling, especially in fields like epidemiology and social sciences, where contact rates, event occurrences, or disease incidences are often expressed as counts. Typically, count data are modeled using Poisson regression, which assumes that the mean and variance of the counts are equal, a property known as equidispersion. However, real-world count data often exhibit overdispersion where variance exceeds the mean or an excess of zeros that do not fit the basic Poisson assumptions. This article, which seeks to model the contact rate matrices of the population in Mayotte, provides a relevant example of these complexities.

To address overdispersion, negative binomial models are often employed, as they introduce an extra parameter to account for greater variance than the mean [6]. Despite their flexibility, these models do not always accommodate the presence of structural zeros cases where zero counts arise from an entirely separate process than the typical count-generating process. Zero-inflated models (ZIMs) were introduced to address this issue, enabling the model to account for zero values resulting from distinct processes. For instance, a count of zero contacts in the dataset might reflect either a real absence of contact (a structural zero) or an occurrence due to randomness in contact behavior.

The Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) models are widely used variants of these approaches [11]. They model the data as a mixture of two components: a point mass at zero and a count distribution (Poisson or negative binomial). For this study, we employ the Zero-Inflated

Bell (ZIBell) model introduced by [12] as it provides advantages in capturing the characteristics of contact data from Mayotte, where zero counts and overdispersion are both prevalent. The ZIBell model extends the Bell distribution, which is itself an alternative count model, particularly effective when modeling overdispersed data with zero inflation. This model assumes two cases for each observation: a structural zero with a given probability $\tau$ and a count observation drawn from a Bell distribution with probability $1-\tau$.

Formally, the probability mass function of the ZIBell random variable can be expressed as follows:

$$Pr(Y=y) = \begin{cases} \tau + (1-\tau)\exp\left(1-e^{W(\mu)}\right) & \text{if } y=0, \\ (1-\tau)\exp\left(1-e^{W(\mu)}\right)\frac{W(\mu)^y}{y!} & \text{if } y>0, \end{cases}$$

where $\mu > 0$ and $\tau \in (0,1)$. Not that $W(\cdot)$ is the Lambert function [**?** ] and $B_y$ is the Bell number defined by (see [5]) The ZIBell model thus accommodates both excess zeros and a varying count structure, making it a robust choice for data with zero inflation and overdispersion.

In addition to model selection, robust parameter estimation is essential for real-world applications where data contamination or outliers may distort results. Traditional Maximum Likelihood Estimation (MLE) may not be sufficient for datasets like those in this study, which may contain erroneous high counts or uncharacteristic patterns. Thus, we apply the Minimum Density Power Divergence Estimator (MDPDE), as developed by Basu et al. [4], to achieve robustness against such anomalies. The MDPDE minimizes a divergence measure between the observed and theoretical density functions, providing estimates that are both efficient and resilient to data contamination [4]. This method ensures consistency and asymptotic normality, key properties for reliable statistical inference in population contact studies.

In summary, the ZIBell model, coupled with MDPDE, is particularly suited for the objectives of this paper. By accurately estimating the contact rate matrices with consideration of excess zeros and outliers, the approach provides a robust and precise analysis of contact behavior in Mayotte. This framework also supports broader applications in fields where data contamination and zero inflation are frequent challenges, demonstrating the versatility and practical relevance of advanced count data models.

## 1.2 Materiel and data

Information on social contacts was obtained using cross sectional surveys conducted by the french regional Agency ARS in the island of Mayotte. The surveys were conducted between october and december 2021 with the oral informed consent of participants. Participants were assigned a random day of the week to record every person they had contact with. Briefly, only one person in each household was asked to participate in the study. Paper diaries were given face to face to participants. Participants were coached on how to fill in the diary. Participants were instructed to record contacted individuals only once in the diary. A contact was defined as either skin-to-skin contact such as a kiss or handshake (a physical contact), or a two-way conversation with three or more words in the physical presence of another person but no skin-to-skin contact (a nonphysical contact). Participants were also asked to provide information about the age and sex of each contact person. If the age of a contact person was not known precisely, participants were asked to provide an estimate of the age range. For each contact, participants were asked to record location (home and outside home) as well as the average number of usual contacts with this individual. The survey sample covered in Mayotte, with quota sampling by age, sex, and commune, randomly selected from population records, excluding persons younger than 1 year of age. Participants were sent a written invitation for a face-to-face interview. If necessary, respondents or their parents were visited at home and approached in another language than French. During the interviews, the participants reported their own age and were asked about the number of persons in their household, excluding themselves, and the number of different persons they conversed with during a typical week, excluding household members. A total of 3,258 responded and completed the questionnaire. During the social contact survey conducted in 2022, a total of 3,258 participants were recruited. During the interviews, the participants reported their own age and were asked about the number of persons in their household as well as their age group, excluding themselves, and the number of different persons they have contact during a typical week day or weekend. The investigation diary categorised people into five age groups: [0,18], [19–24], [25–49], [50-64] and [65+]. During the interviews, the participants reported their own age and were asked about the number of persons in their household, excluding themselves, and the

number of different persons they have contact with (from conversation to a given closed action) during a typical week,day, including household members, specified for the above age group. Diaries recorded basic socio-demographic information about the participant, including employment status, level of completed education, household composition, age, gender; etc...

Note that during the interviews, participants' answers with higher values 900 correspond to special cases. In the sequel and in this study, the training data will concerned values for the number of contacts less than cmax = 50 and values of household size less than hhmax = 100 otherwise they are considered as outliers. Of the 4.467 invited participants (initial data), 2.932 responses excluding incomplete answers (training data) met this criteria (65 percent of the invited participants) for further analysis. In order to compute the population level age contact matrix, we use a formal description. This weight effectively describes how much an ego and its contacts should be considered in order to receive a contact matrix for a closer-to-representative population. The population level contact matrix is computed by randomly selected from population records, excluding persons younger than 1 year of age. For each participant, we added the answers to both questions to obtain the age-specific number of different conversation partners the participant encountered during a typical week. Incomplete or inconsistent answers were excluded firstly from the analysis and secondly estimated using statistical method. Each individual in a population can be categorized using a multitude of attributes (e.g., age, ethnicity, education level, employment, household size,etc). We can distinguish attributes by their range: some take on continuous values (e.g., age), while others are categorical (e.g., education level), discrete-valued or even binary (e.g., vaccinated against COVID-19 or not). As a whole, we can stratify a population based on a selection of attributes. In what follows, we assume all attributes take on only finitely many different values. This does not limit the usability of the developed methods, as binning can turn any continuous-valued attribute such as age into one with finitely-many choices, without losing substantial information if the number of bins is large. We considered both physical and non-physical contacts, including additional professional contacts reported by participants. We modeled the number of contacts using a GAM negative binomial regression model to account for over-dispersion. Sociodemographic characteristics, the health status of participants and microscopic time settings (weekdays/weekends and regular/holiday period), household size, weights, were included as possible determinants (Descriptive statistics see Additional file 1). We performed variable selection using a random forest analysis and the likelihood ratio test. Interactions between age and microscopic time settings were retained, as they were the two most significant determinants of the number of contacts reported in the literature [10]. We use the negative binomial distribution to describe the self-reported number of conversational partners encountered during a typical week in age class i by participants in age class j. Finally, we compared reported household size with social contacts for participants that had at least one reported physical or conversational contact. To explore the relationship between household size, social contacts, we used a generalized additive model [30]. The model was defined as where y was the binary outcome variable (i.e. reported contacts made), x was the predictor (i.e. household size,..), g was the link function, b was the intercept, a was age (to adjust for possible confounding), and f was a smooth function represented by a penalized regression spline. Results for fitted GAMs are shown for a = 30 as example. The contacts reported in several places are assigned according to the following locations and types of contact: home, work, school, transport, leisure and other locations, contact with friends, contact with internal family (home), external family (other homes), contact with neighbors, public places (mosquees, etc) and so on. Variables: sex, age category contatc, employment, household size and age category, type of movement, vaccination status, covid contamination and type of covid contamination, places or types of contamination.

# 2 Methodology

In this section, we briefly recall the definition of the ZIBell model and we describe density power divergence estimator

## 2.1 Zero inflated Bell model

The zero-inflated Bell model is a mixture of two distributions, a degenerate distribution at zero and a Bell distribution, so that for the ith observation $y_i (i = 1, 2, \cdots, n)$, there are two possible cases, the first one is a structural zero with probability $\tau_i$ and the second one is a count data model with probability

$1-\tau_i$. Hence, the probability mass function of the ZIBell random variable can be written as follows:

$$f\left(y_i \mid \tau_i, \mu_i\right)=\begin{cases}\tau_i+\left(1-\tau_i\right)\exp(1-e^{W(\mu_i)}) & \text{if} \quad y_i=0, \\ \\ \left(1-\tau_i\right)\exp(1-e^{W(\mu_i)})\frac{W(\mu_i)^y B_{y_i}}{y_i!} & \text{if} \quad y_i>0,\end{cases} \tag{1}$$

where $\mu_i = e^{\beta^\top \mathbf{X}}$ and $\beta = \left(\beta_1, \beta_2, \cdots, \beta_p\right)^\top$ is the vector of the count parameters and $\mathbf{X}_i = \left(x_{i1}, x_{i2}, \ldots, x_{ip}\right)^\top$. Also, $\tau_i = e^{\gamma^\top \mathbf{Z}}/(1+e^{\gamma^\top \mathbf{Z}}), \quad 0 \leq \tau_i \leq 1, \quad i=1,2,\ldots,n$ where $\gamma = \left(\gamma_1, \gamma_2, \ldots, \gamma_q\right)^\top$ is the parameter vector in the structural zero component and $\mathbf{Z}_i = \left(z_{i1}, z_{i2}, \cdots, z_{iq}\right)^\top$. We denote unknown parameters vector as $\theta = \left(\beta^\top, \gamma^\top\right)^\top$, where $\theta$ is a $k \times 1$ vector, so that $k = p + q$.

We assume that $\left(Y_1, \mathbf{X}_1, \mathbf{Z}_1\right), \left(Y_2, \mathbf{X}_2, \mathbf{Z}_2\right), \ldots, \left(Y_n, \mathbf{X}_n, \mathbf{Z}_n\right)$ are $n$ independent observed vectors from model (2). The likelihood function of the ZIBell model is as:

$$L_{ZIBell}(\theta)=\prod_{i=1}^{n}\left(\tau_i+\left(1-\tau_i\right)\exp(1-e^{W(\mu_i)})\right)^{J_i}\left(1-\tau_i\right)\exp(1-e^{W(\mu_i)})\frac{W(\mu_i)^y B_{y_i}}{y_i!}\Bigg)^{1-J_i}$$

where $J_i = I\left(y_i = 0\right)$ and $\bar{J}_i = 1 - J_i = I\left(y_i > 0\right)$. Therefore, the log likelihood function of the ZIBell model can be written as follows:

$$\begin{aligned}\ell(\theta) = \sum_{i=1}^{n}\Bigg\{ & J_i \log\left[e^{\gamma^\top \mathbf{Z}_i}+\exp(1-e^{W(e^{\beta^\top \mathbf{X}_i})})\right]-\log\left(1+e^{\gamma^\top \mathbf{Z}_i}\right) \\ & +\bar{J}_i\left[Y_i\left(\beta^\top \mathbf{X}_i - W(e^{\beta^\top \mathbf{X}_i})\right) - e^{W(e^{\beta^\top \mathbf{X}_i})} - \log\left(Y_i!\right) + \log B_y + 1\right]\Bigg\}.\end{aligned}$$

The maximum likelihood estimator of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ can be obtained by maximizing $\ell(\theta)$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.

The next section describes the problem and the proposed estimator.

## 2.2 Minimum power divergence estimator

In this subsection, we briefly describe the use of the density power divergence to obtain an estimation of the parameters of the model (1). The asymptotic behavior of the estimated parameter is also studied. Assume that the observations $(Y_1, \mathbf{X}_1, \mathbf{Z}_1), \cdots, (Y_n, \mathbf{X}_n, \mathbf{Z}_n)$ are generated from (**??**) according to the true parameter $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ which is unknown; i.e., $g(. \mid \boldsymbol{\theta}^*)$ is the true conditional density of $Y_i \mid \mathbf{X}_i, \mathbf{Z}_i, i = 1, \cdots, n$. Let $\mathbb{G} = \{g(. \mid \theta), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. be the parametric family of density functions indexed by $\theta \in \Theta$. To estimate $\theta^*$, [4] have proposed a method which consists to choose the "best approximating distribution" of $Y_i \mid \mathbf{X}_i, \mathbf{Z}_i, i = 1, \cdots, n.$ in the family $\mathbb{G}$ by minimizing a divergence $d_\alpha$ between density functions $g(. \mid \boldsymbol{\theta})$ and $g(. \mid \boldsymbol{\theta}^*)$. The density power divergence $d_\alpha$ between two density functions $g$ and $g_*$ is defined by (in the discrete set-up)

$$d_\alpha(g, g_*)=\begin{cases}\sum\limits_{y=0}^{\infty}\left\{g^{1+\alpha}(y)-(1+\frac{1}{\alpha})g_*(y)g^\alpha(y)+\frac{1}{\alpha}g_*^{1+\alpha}(y)\right\}, & \alpha>0 \\ \\ \sum\limits_{y=0}^{\infty}\left\{g_*(y)\left(\log g_*(y)-\log g(y)\right)\right\}, & \alpha=0\end{cases}$$

So, the empirical objective function (based on the divergence between conditional density functions) up to some terms which are independent of is $H_{\alpha,n}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\alpha,i}(\theta)$ where

$$\ell_{\alpha,i}(\theta) = \begin{cases} \sum\limits_{y=0}^{\infty} g^{1+\alpha}(y \mid \theta) - (1+\frac{1}{\alpha}) g^{\alpha}(Y_i \mid \theta), & \alpha > 0 \\ -\log g(Y_i \mid \theta), & \alpha = 0 \end{cases}$$

$$\ell_{\alpha,i}(\theta) = \begin{cases} m_{\alpha,i,1}(\boldsymbol{\theta}) + m_{\alpha,i,2}(\boldsymbol{\theta}) - (1+\frac{1}{\alpha})\{l_{\alpha,i,1}(\boldsymbol{\theta}) I(Y_i = 0) + l_{\alpha,i,1}(\boldsymbol{\theta}) I(Y_i > 0)\}, & \alpha > 0 \\ -\log l_{\alpha,i,1}(\boldsymbol{\theta}) I(Y_i = 0) - \log l_{\alpha,i,2}(\boldsymbol{\theta}) I(Y_i > 0), & \alpha = 0 \end{cases}$$

$$m_{\alpha,i,1} = \left\{\omega_i + (1-\omega_i)\exp\left(1 - e^{W(\mu_i)}\right)\right\}^{1+\alpha} = \left\{\frac{e^{\gamma^\top \mathbf{Z}_i} + \exp\left(1 - e^{W(e^{\beta^\top \mathbf{X}_i})}\right)}{1 + e^{\gamma^\top \mathbf{Z}_i}}\right\}^{1+\alpha},$$

$$m_{\alpha,i,2} = \sum_{y=0}^{\infty}\left\{(1-\omega_i)\exp\left(1 - e^{W(\mu_i)}\right)\frac{W(\mu_i)^y B_y}{y!}\right\}^{1+\alpha} = \sum_{y=0}^{\infty}\left\{\frac{\exp\left(1 - e^{W(e^{\beta^\top \mathbf{X}_i})}\right)}{1 + e^{\gamma^\top \mathbf{Z}_i}}\frac{W(e^{\beta^\top \mathbf{X}_i})^y B_y}{y!}\right\}^{1+\alpha},$$

$$l_{\alpha,i,1}(\boldsymbol{\theta}) = \left\{\omega_i + (1-\omega_i)\exp\left(1 - e^{W(\mu_i)}\right)\right\}^{\alpha} = \left\{\frac{e^{\gamma^\top \mathbf{Z}_i} + \exp\left(1 - e^{W(e^{\beta^\top \mathbf{X}_i})}\right)}{1 + e^{\gamma^\top \mathbf{Z}_i}}\right\}^{\alpha},$$

$$l_{\alpha,i,2}(\boldsymbol{\theta}) = \left\{(1-\omega_i)\exp\left(1 - e^{W(\mu_i)}\right)\frac{W(\mu_i)^y B_y}{y!}\right\}^{\alpha} = \left\{\frac{\exp\left(1 - e^{W(e^{\beta^\top \mathbf{X}_i})}\right)}{1 + e^{\gamma^\top \mathbf{Z}_i}}\frac{W(e^{\beta^\top \mathbf{X}_i})^y B_y}{y!}\right\}^{\alpha}.$$

## 2.3 Asymptotic properties of the MDPDE

To rigorously verify the consistency and asymptotic normality of the Minimum Density Power Divergence Estimator (MDPDE) for the Zero-Inflated Bell (ZIBell) model, we can assume the following conditions, which are commonly used in asymptotic theory for generalized divergence estimators:

- **(C1)** There exists a unique true parameter vector $\theta^* \in \Theta$ such that the density $g(\cdot \mid \theta^*)$ represents the true model.
- **(C2)** The parameter space $\Theta$ is a compact subset of $\mathbb{R}^k$.
- **(C3)** The empirical objective function $H_{\alpha,n}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\alpha,i}(\theta)$ is continuous and differentiable with respect to $\theta$ in a neighborhood of $\theta^*$. Additionally, the first and second derivatives of $H_{\alpha,n}(\theta)$ with respect to $\theta$ exist and are continuous.
- **(C4)** The empirical objective function $H_{\alpha,n}(\theta)$ converges uniformly in probability to the theoretical objective function $H_{\alpha}(\theta)$ as $n \to \infty$, i.e., $\sup_{\theta\in\Theta} |H_{\alpha,n}(\theta) - H_{\alpha}(\theta)| \xrightarrow{p} 0$.
- **(C5)** The score function $S_{\alpha,i}(\theta) = \frac{\partial \ell_{\alpha,i}(\theta)}{\partial \theta}$ evaluated at $\theta = \theta^*$ has finite moments up to a sufficient order (typically the second moment). This implies that $\mathbb{E}[S_{\alpha,i}(\theta^*)] = 0$ and that the variance-covariance matrix of $S_{\alpha,i}(\theta^*)$, denoted by $I_{\alpha}(\theta^*) = \mathbb{E}\left[S_{\alpha,i}(\theta^*) S_{\alpha,i}(\theta^*)^\top\right]$ exists and is positive definite.
- **(C6)** The observations $(Y_i, X_i, Z_i)$ for $i = 1, \ldots, n$ are independent. This assumption simplifies the derivations of asymptotic properties, especially the central limit theorem for establishing asymptotic normality.
- **(C7)** The Fisher information-like matrix $I_{\alpha}(\theta^*) = \mathbb{E}\left[S_{\alpha,i}(\theta^*) S_{\alpha,i}(\theta^*)^\top\right]$ is positive definite. This condition ensures that the estimator's variance-covariance matrix is invertible, allowing for a well-defined asymptotic distribution.

We are now in position to state our first result:

**Theorem 2.1** *Assume that conditions* ***(C1)-(C5)*** *hold. Then, the MDPDE estimator* $\hat{\theta}_\alpha$ *is consistent, that is:*

$$\hat{\theta}_\alpha \xrightarrow{p} \theta^* \quad as\ n \to \infty.$$

**Proof of Theorem 2.1** To prove consistency, we need to show that the MDPDE estimator $\hat{\theta}^\alpha$ converges in probability to the true parameter $\theta^*$ as $n \to \infty$.
Define the empirical mean objective function of the MDPDE estimator as:

$$H_{\alpha,n}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\alpha,i}(\theta),$$

where $\ell_{\alpha,i}(\theta)$ represents the contribution of observation $i$ to the density power divergence criterion.
By condition **(C4)**, we know that this empirical objective function converges uniformly in probability to the theoretical objective function $H_\alpha(\theta)$, i.e.,

$$\sup_{\theta\in\Theta} |H_{\alpha,n}(\theta) - H_\alpha(\theta)| \xrightarrow{p} 0.$$

Condition **(C1)** states that the true parameter $\theta^*$ is the unique minimizer of $H_\alpha(\theta)$ over $\Theta$. By the arg min principle in asymptotic approximation, the estimator $\hat{\theta}^\alpha$, which minimizes $H_{\alpha,n}(\theta)$, thus converges in probability to $\theta^*$.

Thus,

$$\hat{\theta}^\alpha \xrightarrow{p} \theta^* \quad \text{as } n \to \infty,$$

which proves the consistency of $\hat{\theta}^\alpha$.

Asymptotic properties of $\hat{\theta}_n$ are now presented in the following theorems.
**Theorem 2.2** *Assume that conditions* ***(C1)-(C7)*** *hold.Then, The MDPDE estimator* $\hat{\theta}_\alpha$ *follows an asymptotically normal distribution:*

$$\sqrt{n}\left(\hat{\theta}_\alpha - \theta^*\right) \xrightarrow{d} N\left(0, I_\alpha^{-1}(\theta^*)\right) \quad as\ n \to \infty,$$

*where* $I_\alpha(\theta^*)$ *is the asymptotic variance-covariance matrix of* $S_{\alpha,i}(\theta^*)$.
**Proof of Theorem 2.2**. To demonstrate that $\sqrt{n}(\hat{\theta}^\alpha - \theta^*)$ is asymptotically normally distributed, we consider the score function associated with the MDPDE estimator, given by:

$$S_{\alpha,n}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell_{\alpha,i}(\theta)}{\partial\theta}.$$

Performing a Taylor expansion of $S_{\alpha,n}(\theta)$ around $\theta = \theta^*$, we obtain:

$$S_{\alpha,n}(\hat{\theta}^\alpha) = S_{\alpha,n}(\theta^*) + J_{\alpha,n}(\theta^*)\left(\hat{\theta}^\alpha - \theta^*\right) + o_p\left(\|\hat{\theta}^\alpha - \theta^*\|\right),$$

where

$$J_{\alpha,n}(\theta^*) = \left.\frac{\partial S_{\alpha,n}(\theta)}{\partial\theta}\right|_{\theta=\theta^*}$$

represents the Jacobian matrix of $S_{\alpha,n}(\theta)$ at $\theta^*$.
Since $\hat{\theta}^\alpha$ satisfies the score equation $S_{\alpha,n}(\hat{\theta}^\alpha) = 0$, we deduce that:

$$0 = S_{\alpha,n}(\theta^*) + J_{\alpha,n}(\theta^*)\left(\hat{\theta}^\alpha - \theta^*\right) + o_p\left(\|\hat{\theta}^\alpha - \theta^*\|\right).$$

Under condition **(C3)**, we know that $J_{\alpha,n}(\theta^*)$ converges in probability to the asymptotic Fisher information matrix $J_\alpha(\theta^*)$, that is: $J_{\alpha,n}(\theta^*) \xrightarrow{p} J_\alpha(\theta^*)$.
According to condition **(C5)**, the score function $S_{\alpha,n}(\theta^*)$ satisfies the central limit theorem. Thus, we have: $\sqrt{n}S_{\alpha,n}(\theta^*) \xrightarrow{d} \mathcal{N}(0, I_\alpha(\theta^*))$, where

$$I_\alpha(\theta^*) = E[S_{\alpha,i}(\theta^*)S_{\alpha,i}(\theta^*)^T]$$

is the covariance matrix of $S_{\alpha,i}(\theta^*)$.

Solving for $\hat{\theta}^\alpha - \theta^*$ in the score equation, we have:

$$\sqrt{n}(\hat{\theta}^\alpha - \theta^*) = -J_\alpha(\theta^*)^{-1}\sqrt{n}S_{\alpha,n}(\theta^*) + o_p(1).$$

Since $\sqrt{n}S_{\alpha,n}(\theta^*)$ is asymptotically normal, it follows that:

$$\sqrt{n}(\hat{\theta}^\alpha - \theta^*) \xrightarrow{d} \mathcal{N}(0, J_\alpha(\theta^*)^{-1}I_\alpha(\theta^*)J_\alpha(\theta^*)^{-1}).$$

Defining the asymptotic covariance matrix of the estimator as:

$$V_\alpha(\theta^*) = J_\alpha(\theta^*)^{-1}I_\alpha(\theta^*)J_\alpha(\theta^*)^{-1},$$

we conclude that the sought asymptotic distribution is:

$$\sqrt{n}(\hat{\theta}^\alpha - \theta^*) \xrightarrow{d} \mathcal{N}(0, V_\alpha(\theta^*)).$$

Thus, the MDPDE estimator $\hat{\theta}^\alpha$ is asymptotically normal with a covariance matrix given by $V_\alpha(\theta^*) = J_\alpha(\theta^*)^{-1}I_\alpha(\theta^*)J_\alpha(\theta^*)^{-1}$.

# 3 Numerical Study

# 4 Numerical Study

In this section, we examine the performance of the Minimum Density Power Divergence Estimator (MDPDE) for finite sample sizes. The robustness and efficiency of the MDPDE are evaluated and compared to those of the Maximum Likelihood Estimator (MLE), which corresponds to the special case of the MDPDE with $\alpha = 0$. In particular, we are interested in the stability of these estimators when the data is contaminated with outliers.

## 4.1 Simulation Setup

We simulate data from a Zero-Inflated Bell (ZI-Bell) regression model ((2.1)-(2.2)-(2.3)), defined by:

$$\log(\mu_i) = \beta_1 + \beta_2 X_{i2} \quad \text{and} \quad \text{logit}(\pi_i) = \gamma Z_{i1},$$

where $X_{i1} = Z_{i1} = 1$, and the covariate $X_{i2}$ is drawn from a standard normal distribution $\mathcal{N}(0,1)$. The parameters $\beta_1$ and $\beta_2$ are set to $\beta_1 = (-0.5, 0.5, -0.75, -1, 0)^\top$ and $\beta_2 = 1.2$. We consider two values for $\gamma$, specifically $\gamma = -1.1$ and $\gamma = 0.2$. With these values, the average proportion $c$ of non-inflated data in the simulated datasets is 0.25 and 0.50, respectively. To evaluate the performance of the estimators, we conduct Monte Carlo simulations with $N = 500$ repetitions for sample sizes of $n = 500$ and $n = 1000$. The data generation process involves two scenarios:
**Scenario 1: No contamination**. All data points follow the ZI-Bell model without any outliers.
**Scenario 2: Contaminated data**. A fraction of the data is intentionally perturbed by introducing outliers to assess the robustness of the estimators. Suppose the contaminated process $Y_{c,i}$ is observed as $Y_{c,i} = Y_i + P_i Y_{0,i}$, where $Y_i$ is generated from (1), $P_i$ is an i.i.d. Bernoulli random variable with a success probability $p$, and $Y_{0,i}$ is an i.i.d. Poisson random variable with mean $\mu$. The variables $P_i$, $Y_{0,i}$, and $Y_i$ are assumed to be independent. For $p = 0.008$ and $\mu = 10$, the corresponding results are summarized in Table 2.

Based on the $N$ estimates, we compute, for each simulation scenario, the sample mean, empirical bias, and RMSE of the estimates $\hat{\beta}_{j,n}$ and $\hat{\gamma}_{k,n}$ over the $N$ simulated samples. For instance, the empirical bias of $\hat{\beta}_{j,n}$ is defined as follows: $\bar{\beta}_{1,1} - \beta_{1,1}$, where $\hat{\beta}_{1,1}^{(k)}$ is the power divergence estimate of the first component $\beta_{1,1}$ of $\beta_1$, obtained from the $k$-th simulated sample ($k = 1, \ldots, N$), and $\bar{\beta}_{1,1} := \frac{1}{N}\sum_{k=1}^{N} \hat{\beta}_{1,1}^{(k)}$. We use two criteria to evaluate the performance of the estimators: the sample mean, which measures the bias of the estimator by comparing the estimated parameters to the true values, and the Root Mean Square Error (RMSE), which quantifies the overall accuracy of the estimators by accounting for both bias and variance. In Tables 1 to 2, the symbol † denotes the minimal RMSEs, and shaded areas indicate cases where the MDPDE achieves lower RMSEs than the MLE.

## 4.2 Results

Tables 1 and 2 present the parameter estimation performance for the Zero-Inflated Bell (ZI-Bell) model under different values of the robustness parameter $\alpha$ and different sample sizes ($n = 500$ and $n = 1000$). The tables show variations in performance depending on whether the data contains outliers.

### *Scenario 1: Data without Contamination*

Table 1 displays the simulation results for the scenario without contamination. Each cell in the table shows the mean parameter estimates, with the bias and RMSE indicated in parentheses. The analysis is conducted for various values of the tuning parameter $\alpha$ ($\alpha = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 1$). The results indicate that the estimators perform well in the absence of contamination. The estimated value of each parameter gradually approaches the true values as $\alpha$ increases, demonstrating improved robustness to minor perturbations in the data. For $\alpha = 0$ (MLE), the bias is minimal, and the RMSE decreases as the sample size increases from $n = 500$ to $n = 1000$, indicating that larger sample sizes improve the accuracy of the estimates. As $\alpha$ increases, the estimators show a slight increase in bias and RMSE, which is expected due to the downweighting of extreme observations characteristic of the MDPDE. Moreover, the results show that in the absence of outliers, the MLE has minimal RMSEs for all parameters, except for the parameter $\gamma$, where the MDPDE with $\alpha = 0.2$ achieves the lowest RMSE.

### *Scenario 2: Data with Contamination*

Table 2 presents estimates for the case where the data is contaminated with outliers, allowing an assessment of the robustness of MDPDE estimates under perturbed data conditions. For $\alpha = 0$, the MLE shows significant increases in bias and RMSE for all estimates, especially for $\gamma$, indicating sensitivity to data contamination. In contrast, as $\alpha$ increases, the MDPDE becomes more robust, with a marked reduction in bias and RMSE compared to the MLE, particularly for higher values of $\alpha$. For example, for $\alpha = 0.2$ and $n = 1000$, the MDPDE estimates exhibit significantly lower biases and RMSEs than those of the MLE, suggesting better robustness against outliers. This trend persists as $\alpha$ increases, confirming that higher values of $\alpha$ improve resistance to contamination at the cost of a slight increase in bias in the absence of contamination.

Regarding the effect of $\alpha$ on estimates, it is noted that as $\alpha$ increases, biases gradually decrease for all estimates, indicating better robustness to contaminated data. Bias and RMSE values improve considerably for $\alpha$ values between 0.1 and 0.3, which seems to be an ideal range for this type of data. Moreover, the results show that the proportion of zero-inflation impacts biases: for $\gamma$, the presence of zero-inflation leads to higher biases when $\alpha$ is low (notably for $\alpha = 0$), but these biases decrease as $\alpha$ increases, confirming that the MDPDE offers better robustness for handling zero-inflation cases. These simulation results highlight a trade-off between efficiency and robustness for different values of $\alpha$. The MLE ($\alpha = 0$) performs well in clean data scenarios but is sensitive to contamination. In contrast, the MDPDE ($\alpha > 0$) provides more robust estimates at the expense of slight efficiency losses for uncontaminated data, demonstrating its suitability for practical applications where data contamination is a concern.

**Table 1**: Simulation results for the case without outliers in the data. Sample mean(biais/RMSE) estimators

| | | $n = 500$ | | | $n = 1000$ | | |
|---|---|---|---|---|---|---|---|
| 4-79-11 $c$ | $\alpha$ | $\widehat{\beta}_0$ | $\widehat{\beta}_1$ | $\widehat{\gamma}$ | $\hat{\beta}_0$ | $\hat{\beta}_1$ | $\widehat{\gamma}$ |
| .25 | | | | | | | |
| | 0 | -0.503(-0.003/0.113)$^{\dagger}$ | 1.201(0.001/0.073)$^{\dagger}$ | -1.147(-0.047/0.297)$^{\dagger}$ | -0.498(0.002/0.077)$^{\dagger}$ | 1.200(0.000/0.048)$^{\dagger}$ | -1.099(0.001/0.191)$^{\dagger}$ |
| | 0.1 | -0.510(-0.010/0.115) | 1.219(0.019/0.079) | -1.159(-0.059/0.299) | -0.506 (-0.006/0.078) | 1.219(0.019/0.053) | -1.111(-0.011/0.191) |
| | 0.2 | -0.509(-0.009/0.117) | 1.217(0.017/0.081) | -1.155(-0.055/0.299) | -0.505(-0.005/0.080) | 1.217(0.017/0.054) | -1.108(-0.008/ 0.191) |
| | 0.3 | -0.508(-0.008/0.120 ) | 1.215(0.015/0.083) | -1.153(-0.053/0.300) | -0.503(-0.003/0.082) | 1.215(0.015/0.055) | -1.105(-0.005/0.191) |
| | 0.4 | -0.507(-0.007/0.125) | 1.214(0.014/0.086) | -1.152(-0.052/0.302) | -0.502(-0.002/0.086) | 1.213(0.013/0.057) | -1.104(-0.004/0.192) |
| | 0.5 | -0.506(-0.006/0.129 ) | 1.212(0.012/0.089) | -1.152( -0.052/0.303) | -0.502(-0.002/0.089) | 1.212(0.012/0.059) | -1.102(-0.002/0.193) |
| | 0.75 | -0.506(-0.006/0.137) | 1.210( 0.010/0.097) | -1.152(-0.052/0.308) | -0.500(0.000/0.097) | 1.209(0.009/0.066 ) | -1.101(-0.001/0.196) |
| | 1 | -0.506(-0.006/0.142) | 1.209(0.009/0.105) | -1.153(-0.053/0.312) | -0.500(0.000/0.102) | 1.207(0.007/0.073) | -1.101(-0.001/0.199) |
| .50 | | | | | | | |
| | 0 | -0.499(0.001/0.149)$^{\dagger}$ | 1.197(-0.003/0.089)$^{\dagger}$ | 0.193(-0.007/0.183)$^{\dagger}$ | -0.510(-0.010/0.09)$^{\dagger}$ | 1.202(0.002/0.066)$^{\dagger}$ | 0.2023(0.003/0.133) |
| | 0.1 | -0.508(-0.008/0.153) | 1.216(0.016/0.096) | 0.184(-0.016/0.185) | -0.520(-0.020/0.093) | 1.222(0.022/0.074) | 0.193( -0.007/0.132) |
| | 0.2 | -0.505(-0.005/0.159) | 1.213(0.013/ 0.098) | 0.186(-0.014/0.186) | -0.518(-0.018/0.095) | 1.219(0.019/0.075) | 0.195(-0.005/ 0.131)$^{\dagger}$ |
| | 0.3 | -0.504(-0.004/0.165) | 1.211(0.011/0.101) | 0.188(-0.012/ 0.188) | -0.517(-0.017/0.099) | 1.217(0.017/0.076 ) | 0.197(-0.003/0.132) |
| | 0.4 | -0.503(-0.003/0.172) | 1.209(0.009/0.105) | 0.189(-0.011/ 0.190) | -0.516(-0.016/0.103) | 1.216( 0.016/0.079 ) | 0.198(-0.002/0.133) |
| | 0.5 | -0.503(-0.003/0.179) | 1.207(0.007/0.111) | 0.189(-0.011/0.193) | -0.516(-0.016/0.108) | 1.215(0.015/0.081) | 0.199(-0.001/0.134) |
| | 0.75 | -0.506(-0.006/0.197) | 1.205(0.005/0.128) | 0.188(-0.012/0.199) | -0.518(-0.018/0.120) | 1.215(0.015/0.090) | 0.198(-0.002/0.138) |
| | 1 | -0.510(-0.010/0.212) | 1.205(0.005/ 0.147) | 0.186(-0.014/0.206) | -0.521(-0.021/0.129) | 1.219(0.019/ 0.101) | 0.197(-0.003/0.143) |

**Table 2**: Simulation results for the the case in which the data are contaminated by outliers. Sample mean(biais /RMSE) estimators

| | | $n = 500$ | | | $n = 1000$ | | |
|---|---|---|---|---|---|---|---|
| 4-79-11 $c$ | $\alpha$ | $\widehat{\beta}_0$ | $\widehat{\beta}_1$ | $\widehat{\gamma}$ | $\hat{\beta}_0$ | $\hat{\beta}_1$ | $\widehat{\gamma}$ |
| .25 | | | | | | | |
| | 0 | -0.428(0.071/0.157) | 1.170(-0.029/0.088) | -1.034(0.065/0.296) | -0.431(0.068/0.117) | 1.175(-0.024/0.062) | -1.008(0.091/0.203) |
| | 0.1 | -0.492(0.008/0.114)$^{\dagger}$ | 1.213(0.013/0.076)$^{\dagger}$ | -1.113(-0.013/0.276)$^{\dagger}$ | 0.488(0.012/0.085)$^{\dagger}$ | 1.215(0.015/0.056)$^{\dagger}$ | -1.080(0.020/0.184)$^{\dagger}$ |
| | 0.2 | -0.504(-0.004/0.115) | 1.216( 0.016/0.078) | -1.128(-0.028/0.281) | -0.449(0.051/0.105) | 1.192(-0.008/0.064 ) | -1.132(-0.032/0.209) |
| | 0.3 | -0.506(-0.006/0.119) | 1.215(0.015/0.081) | -1.132( -0.032/0.285) | -0.468(0.032/0.100) | 1.197(-0.003/0.067) | -1.157(-0.057/0.218) |
| | 0.4 | -0.507(-0.007/0.124) | 1.213(0.013/0.085) | -1.134(-0.034/0.288) | -0.471(0.029/0.103) | 1.195(-0.005/0.071 ) | -1.165(-0.065/0.222) |
| | 0.5 | -0.506(-0.006/0.128) | 1.212(0.012/0.089) | -1.135(-0.035/0.291) | -0.469(0.031/0.107) | 1.191(-0.009/0.075) | -1.168(-0.068/0.225) |
| | 0.75 | -0.503(-0.003/0.142) | 1.208(0.008/ 0.113) | -1.138(-0.038/0.302) | -0.458(0.042/0.116) | 1.178(-0.022/0.087) | -1.168(-0.068/0.228) |
| | 1 | -0.505(-0.002/0.122) | 1.206(0.009/0.117) | -1.127(-0.036/0.305) | -0.447(0.053/0.123) | 1.166(-0.034/0.097) | -1.166(-0.066/0.229) |
| .50 | | | | | | | |
| | 0 | -0.420(0.079/0.210) | 1.166(-0.033/0.123) | 0.249(0.049/0.225) | -0.421(0.078/0.157) | 1.172(-0.027/0.085) | 0.248(0.048/0.150) |
| | 0.1 | -0.485(0.015/0.165)$^{\dagger}$ | 1.211(0.011/ 0.108)$^{\dagger}$ | 0.202(0.002/ 0.209)$^{\dagger}$ | -0.496(0.004/0.113)$^{\dagger}$ | 1.221(0.021/ 0.077)$^{\dagger}$ | 0.210(0.010/0.133) |
| | 0.2 | -0.499(0.001/0.166) | 1.214(0.014/0.109) | 0.194(-0.006/0.209) | -0.510(-0.010/0.114) | 1.225(0.025/0.080) | 0.203 (0.003/ 0.132)$^{\dagger}$ |
| | 0.3 | -0.503(-0.003/0.171) | 1.213(0.013/0.111) | 0.192(-0.008/0.210) | -0.515(-0.015/0.118) | 1.225 (0.025/0.082) | 0.201 (0.001/0.133) |
| | 0.4 | -0.505(-0.005/0.177) | 1.211(0.011/0.116) | 0.191(-0.009/ 0.211) | -0.502(-0.002/0.116) | 1.219(0.019/0.081) | 0.195(-0.005/0.136) |
| | 0.5 | -0.506(-0.006/0.184) | 1.209(0.009/0.121) | 0.190(1.290/1.308) | -0.502(-0.002/0.121) | 1.217(0.017/0.084) | 0.195(-0.005/ 0.138) |
| | 0.75 | -0.508(-0.008/0.203) | 1.207(0.007/0.137) | 0.188(-0.012/0.217) | -0.500(0.000/0.134) | 1.215 (0.015/0.094) | 0.195(-0.005/0.142) |
| | 1 | -0.511(-0.011/0.220) | 1.209(0.009/0.156) | 0.186(-0.014/0.223) | -0.516(-0.016/0.154) | 1.216(0.016/0.109) | 0.198(-0.002/0.145) |

# References


[1] Ali, E., Diop, M. L., & Diop, A. Statistical Inference in a Zero-Inflated Bell Regression Model. *Mathematical Methods of Statistics*, 31(3), 91-104, 2022.

[2] Ali, E., & Pho, K. H. A novel model for count data: zero-inflated Probit Bell model with applications. *Communications in Statistics - Simulation and Computation.*2024. Available from: https://doi.org/10.1080/03610918.2024.2384574

[3] Amin, M., Akram, M. N., & Majid, A. On the estimation of Bell regression model using ridge estimator. *Communications in Statistics - Simulation and Computation*, 52(3), 854-867,2021.

[4] Basu, A., et al. Minimum Density Power Divergence Estimation. *The Annals of Statistics*,(26)40-61,1998.

[5] Bell,E.T.,. Exponential polynomials. *Annal. Math.* (35)258–277, 1934a.

[6] A. C. Cameron and P. K. Trivedi. Regression Analysis of Count Data. *Cambridge University Press*, 2013.

[7] Corless, R.M., Gonnet, G.H.,Hare, D.E.G., Jeffrey, D. and Knuth,D.E. On the LambertW function, *Adv. Comput. Math.* (5)329–359, 1996.

[8] Goeman, J. J., Meijer, R. J., & Chaturvedi, N. Penalized estimation methods for zero-inflated regression models. *Statistical Modelling*, 14(3), 215-237, 2014.

[9] Gibbons, D. G. (1981). A Simulation Study of Some Ridge Estimators. *Journal of the American Statistical Association*, 76(373), 131?139.

[10] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67, 1970

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

[12] Lemonte, A. J., Moreno-Arenas, G., & Castellares, F. Zero-inflated Bell regression models for count data. *Journal of Applied Statistics*, 47(2), 265-286, 2019.

[13] Le Cessie, S., & Van Houwelingen, J. C. Ridge estimators in logistic regression. *Applied Statistics*, 41(1), 191-201, 1992.

[14] Montgomery, D. C., Peck, E. A., & Vining, G. G. Introduction to Linear Regression Analysis. *Wiley*, 2012.

[15] Zou, H., & Hastie, T. Regularization and variable selection via the elastic net. *[*Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320, 2005.

[16] Zakariya Yahya Algamal & Adewale F. Lukman & Mohamed R. Abonazel & Fuad A. Awwad & Niansheng Tang. Performance of the Ridge and Liu Estimators in the zero-inflated Bell Regression Model. *Journal of Mathematics*, Hindawi, vol. 2022, pages 1-15

Cameron, A.C. and Trivedi, P.K. Microeconometrics: Methodsand Applications. *Cambridge: New York*(2005).