Reparametrization of COM-Poisson Regression Models with Applications in the Analysis of Experimental Data

Eduardo E. Ribeiro Jr [1], **Walmes M. Zeviani** [2], **Wagner H. Bonat** [2], **Clarice G. B. Demétrio** [1] **and John Hinde** [3]

[1] Department of Exact Sciences, University of São Paulo - ESALQ, Piracicaba, SP, Brazil
[2] Department of Statistics, Paraná Federal University, Curitiba, PR, Brazil
[3] School of Mathematics, Statistics and Applied Mathematics, National University of Ireland Galway, Galway, Ireland

Address for correspondence: Eduardo E. Ribeiro Jr, Department of Exact Sciences, Luiz de Queiroz College of Agriculture - ESALQ, University of São Paulo - USP, Piracicaba, São Paulo, Pádua Dias, 11, Avenue, CEP–13.418-900 Brazil.
E-mail: `jreduardo@usp.br`.
Phone: (+55) 41 9 8711 9034.
Fax: (+55) 19 3447 6021.

Abstract: The COM-Poisson distribution is a two-parameter generalization of the Poisson distribution that can deal with under-, equi- and overdispersed count data. Unfortunately, its location parameter does not correspond to the expectation, which complicates the interpretation of regression models specified using this distribution. In this paper, we propose a straightforward reparametrization of the COM-Poisson distribution based on an approximation to the expectation. The main advantage of our new parametrization is the interpretation of the regression coefficients in terms of the (approximate) expectation of the count response variable, as is usual in the context of generalized linear models. Estimation and inference for the reparametrized COM-Poisson regression models can be done using the likelihood paradigm. We carried out simulation studies to verify the finite sample properties of the maximum likelihood estimators. The results from our simulation study show that the maximum likelihood estimators are unbiased and consistent for both regression and dispersion parameters. The empirical correlation between the regression and dispersion parameter estimators and nature of the deviance surfaces suggests that these parameters are also approximately orthogonal, which is advantageous for interpretation, inference, and computational efficiency. We illustrate the application of the proposed model through the analysis of three data sets with over-, under- and equidispersed count data. The study of the distribution's properties, through a consideration of dispersion, zero-inflation and heavy tail indexes, together with the results of data analysis, show the flexibility over standard approaches. The computational routines for fitting the original and reparameterized versions of the COM-Poisson regression model and data sets are available in the supplementary material.

Key words: COM-Poisson, Count data, Likelihood inference, Overdispersion, Underdispersion.

1 Introduction

Count data arise from random variables that take non-negative integer values and typically represent the number of times an event occurs in an observation period or region. This kind of data is also common in crop sciences, examples including the number of grains produced by a plant, the number of fruits on a tree, the number of insects captured by a trap, etc. Since the seminal paper of Nelder and Wedderburn (1972), where the class of the generalized linear models (GLMs) was introduced, the analysis of count data often uses a Poisson regression model. This model provides a suitable strategy for the analysis of count data and an efficient Newton scoring algorithm that can be used for fitting.

In spite of the advantages of the Poisson regression model, the Poisson distribution has only one parameter that represents both the expectation and variance of the count random variable. This equality restriction on the relationship between the expectation and variance of the Poisson distribution is referred as equidispersion. However, in practical data analysis such restriction can be unsuitable, since the observed data can present variance both smaller or larger than the mean, leading to under- and overdispersion, respectively. While applying the Poisson regression model to non-equidispersed count data can give consistent parameter estimates, the associated standard errors are inconsistently estimated, which in turn can lead to misleading inferences (Winkelmann, 1995; Bonat et al., 2017).

In practice, overdispersion is widely reported in the literature and may occur due to the absence of relevant covariates, heterogeneity of sampling units, different observational periods/regions not being considered in the analysis, and excess zeros (Hinde and Demétrio, 1998). The case of underdispersion is less often reported in the literature, however, it has been of increasing interest in the statistical community. The processes that reduce the variability are not as well-known as those leading to extra variability. For this reason, there are few models to deal with underdispersed count data. Possible explanatory mechanisms leading to underdispersion may be related to the underlying stochastic process generating the count data. For example, when the time between events is not exponentially distributed, the number of events can be over or underdispersed; a process that motivated the class of duration dependence models (Winkelmann, 1995). Another possible explanation for underdispersion is when the responses correspond to order statistics of component observations, such as maxima of Poisson distributed counts (Steutel and Thiemann, 1989).

Strategies for constructing alternative count distributions are related to the causes of the non-equidispersion. Specifically for overdispersion, Poisson mixture (compound) models are widely applied. One popular example of this approach is the negative-binomial model, where the expectation of the Poisson distribution is assumed to be gamma distributed. However, other distributions can also be used to represent this additional random variation. For example the Poisson-Tweedie model (Bonat et al., 2017) and its special cases the Poisson inverse-Gaussian and Neyman-Type A models assume that the random effects are Tweedie, inverse Gaussian or Poisson distributed, respectively. The Gamma-Count distribution assumes a gamma distribution for the time between events and it can deal with underdispersed as well as overdispersed count data (Zeviani et al., 2014). The subject of this paper, the COM-Poisson distribution, is obtained by a generalization of the Poisson distribution that allows for a non-linear decrease in the ratios of successive probabilities (Shmueli et al., 2005).

The COM-Poisson distribution includes the Poisson and geometric distributions as special cases, as well as the Bernoulli distribution as a limiting case. It can deal with both under- and overdispersed count data. Some recent applications of the COM-Poisson distribution include Lord et al. (2010) for the analysis of traffic crash data, Sellers and Shmueli (2010) for the modelling of airfreight breakage and book purchases, Huang (2017) on the analysis of attendance data, takeover bids and cotton boll counts, and Chatla and Shmueli (2018) to model counts from bike sharing systems. Theoretical results and approximations derived for this distribution are discussed by Shmueli et al. (2005), Daly and Gaunt (2016) and Gaunt et al. (2017). The main disadvantage of the COM-Poisson regression model as presented in Sellers and Shmueli (2010) is that its location parameter does not correspond to the expectation of the distribution. This complicates the interpretation of regression models and means that they are not comparable with standard approaches, such as the Poisson and negative binomial regression models. In order to avoid this issue, Huang (2017) proposed a mean-parametrization of the COM-Poisson distribution. In his approach the mean parameter is obtained as the solution of a non-linear equation defined as an infinite sum, which is computationally demanding and liable to numerical problems.

The main goal of this article is to propose a novel COM-Poisson parametrization based on the mean approximation presented by Shmueli et al. (2005). In this parametrization, the probability mass function is written in terms of μ and ϕ, where μ is the approximate expectation and ϕ is a dispersion parameter. In contrast to the original parametrization, the proposed parametrization leads to regression coefficients directly associated (approximately) with the expectation of the response variable, as is usual in the context of generalized linear models. Consequently, the resulting regression coefficients are comparable to those obtained by standard approaches, such as the Poisson and negative binomial regression models. Furthermore, our novel COM-Poisson parametrization is computationally simpler than the strategy proposed by Huang (2017), since it does not require any numerical method for solving non-linear equations. Also, we show that attractive properties like the orthogonality between dispersion and regression parameters and consistency and asymptotic normality of the maximum likelihood estimators are retained.

This paper is organized as follows. In section 2 we present the COM-Poisson distribution and the approach proposed by Huang (2017). The newly proposed reparametrization is considered in section 3, along with assessment of the moment approximation, and a study of distribution's properties. In section 4 we present estimation and inference for the novel COM-Poisson regression model in a likelihood framework. The properties of the maximum likelihood estimators and their approximate orthogonality are assessed in section 5 through simulation studies. We illustrate the application of the new COM-Poisson regression model with the analysis of three data sets. We provide an R implementation for the COM-Poisson and reparameterized COM-Poisson regression models, together with the analyzed data sets, in the supplementary material.[1].

2 Background

The COM-Poisson distribution generalizes the Poisson distribution in terms of the ratio between the probabilities of two consecutive events by adding an extra dispersion parameter (Sellers and Shmueli, 2010). Let Y be a COM-Poisson random variable, then

$$\frac{\Pr(Y = y - 1)}{\Pr(Y = y)} = \frac{y^\nu}{\lambda}$$

while for the Poisson distribution this ratio is $\frac{y}{\lambda}$ corresponding to $\nu = 1$. This allows the COM-Poisson distribution to deal with non-equidispersed count data. The probability mass function of the COM-Poisson distribution is given by

$$\Pr(Y = y \mid \lambda, \nu) = \frac{\lambda^y}{(y!)^\nu Z(\lambda, \nu)}, \qquad y = 0, 1, 2, \dots, \tag{2.1}$$

where $\lambda > 0$, $\nu \geq 0$ and $Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}$ is a normalizing constant that depends on both parameters.

The $Z(\lambda, \nu)$ series diverges theoretically only when $\nu = 0$ and $\lambda \geq 1$, but numerically, for small values of ν combined with large values of λ, the sum is so huge it results in overflow. Table 1 shows the values of the normalizing constant based on one thousand terms in the summation, that is $\sum_{j=0}^{1000} \lambda^j/(j!)^\nu$, for different values of λ and ϕ.

In the first line of Table 1 we have mathematically divergent series, because $\sum_{j=0}^{\infty} \lambda^j$ is divergent when $\lambda \geq 1$. In other cases the series diverges numerically, due to the computational storage limitation. For both forms of divergence it is impossible to compute COM-Poisson probabilities, therefore, this places a restriction on the feasible parameter space.

An undesirable feature of the COM-Poisson distribution is that the moments cannot be obtained in closed form. Shmueli et al. (2005) and Sellers and Shmueli (2010) using an asymptotic approximation for $Z(\lambda, \nu)$, showed that the expectation and variance of the COM-Poisson distribution can be approximated by

$$\mathrm{E}(Y) \approx \lambda^{1/\nu} - \frac{\nu - 1}{2\nu} \qquad \text{and} \qquad \mathrm{Var}(Y) \approx \frac{\lambda^{1/\nu}}{\nu}, \tag{2.2}$$

with greatest accuracy for $\nu \leq 1$ or $\lambda > 10$. The authors also argue that the mean-variance relationship can be approximated as $\mathrm{Var}(Y) \approx \frac{1}{\nu}\mathrm{E}(Y)$. In section 3, we assess the accuracy of these approximations.

[1] Available on http://www.leg.ufpr.br/~eduardojr/papercompanions .

Table 1: Values for $Z(\lambda, \nu)$ normalizing constant (computed numerically with 1000 terms in the summation) for values of λ (0.5 to 50) and ϕ (0 to 1)

	$\boldsymbol{\lambda}$					
$\boldsymbol{\nu}$	0.5	1	5	10	30	50
0	2.00	divergent*	divergent*	divergent*	divergent*	divergent*
0.1	1.92	7.64	divergent**	divergent**	divergent**	divergent**
0.2	1.86	5.25	3.17e+273	divergent**	divergent**	divergent**
0.3	1.81	4.32	1.60e+29	2.54e+282	divergent**	divergent**
0.4	1.77	3.80	4.71e+10	1.33e+56	divergent**	divergent**
0.5	1.74	3.47	1.34e+06	3.67e+22	3.32e+196	divergent**
0.6	1.72	3.23	2.05e+04	4.99e+12	1.73e+76	4.63e+177
0.7	1.70	3.06	2.37e+03	3.69e+08	4.93e+39	6.93e+81
0.8	1.68	2.92	6.49e+02	2.70e+06	5.09e+24	3.43e+46
0.9	1.66	2.81	2.74e+02	1.47e+05	1.80e+17	2.19e+30
1	1.65	2.72	1.48e+02	2.20e+04	1.07e+13	5.18e+21

divergent* denotes a mathematically divergent series; and divergent** a numerically divergent series.

The COM-Poisson regression model was proposed by Sellers and Shmueli (2010), using this original parametrization. Specifically, the COM-Poisson regression model is defined as $\log(\lambda_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ and the relationship between $\mathrm{E}(Y_i)$ and $\boldsymbol{x}_i$ is modelled indirectly. Huang (2017) shows how to model directly the expectation of the COM-Poisson distribution in a suitable reparametrization. From Equation 2.1, Huang notes that the parameter λ is given, as a function of μ and ν, by the solution to

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

This allows Huang to define a mean-parametrized COM-Poisson regression model with $\log(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ giving a direct relationship between μ and the covariates $\boldsymbol{x}$. In this article, we propose an alternative approximate mean-parametrization of the COM-Poisson distribution in order to avoid the limitations of the original parametrization and the numerical complexity of the Huang's approach.

3 Reparametrized COM-Poisson regression model

The proposed reparametrization of COM-Poisson models is based on the mean approximation (Equation 2.2). We introduce a new parameter μ, using this approximation,

$$\mu = h_\nu(\lambda) = \lambda^{1/\nu} - \frac{\nu - 1}{2\nu} \quad \Rightarrow \quad \lambda = h_\nu^{-1}(\mu) = \left(\mu + \frac{(\nu - 1)}{2\nu}\right)^\nu . \tag{3.1}$$

The dispersion parameter is taken on the log scale for computational convenience, thus $\phi = \log(\nu)$, $\phi \in \mathbb{R}$. The interpretation of ϕ is the same as the ν, but simply on another scale. For $\phi < 0$ and $\phi > 0$ we have the overdispersed and underdispersed cases, respectively. When $\phi = 0$ we have the Poisson distribution as a special case.

In order to assess the accuracy of the moment approximations (Equation 2.2), Figure 1 presents the errors for (a) expectation ($\mu - \mathrm{E}(Y)$) and (b) variance ($\mu \exp(-\phi) - \mathrm{Var}(Y)$). In both cases $\mathrm{E}(Y)$ and $\mathrm{Var}(Y)$ were computed numerically using 500 terms[2]. The dotted lines represent the border between the regions $\nu \leq 1$ and $\lambda > 10^\nu$, in the μ and ϕ scale.

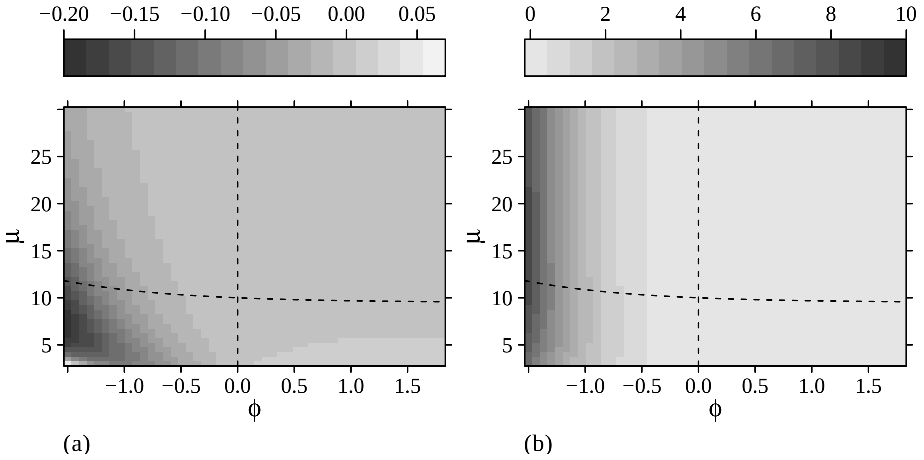

Figure 1: Errors for the approximation of the (a) expectation and (b) variance. Dotted lines represent the restriction for suitable approximations given by Shmueli et al. (2005).

The results in Figure 1 show that the mean approximation is accurate, the largest (absolute) error, for the parameter values evaluated, is -0.197. For the variance approximation, the largest error was 8.4 and it occurs for negative values of ϕ. Interestingly, the errors for the variance are larger for negative values of ϕ and present no clear relation with μ, as opposed to the regions gives by Shmueli et al. (2005) ($\phi \leq 0$ and $\mu > 10 - (\exp(\phi) - 1)/(2\exp(\phi))$).

The results presented in Figure 1(a) support the proposed reparametrization. Replacing λ and ν as functions of μ and ϕ in Equation 2.1, the reparametrized distribution takes the form

$$\Pr(Y = y \mid \mu, \phi) = \left(\mu + \frac{e^\phi - 1}{2e^\phi}\right)^{ye^\phi} \frac{(y!)^{-e^\phi}}{Z(\mu, \phi)}, \qquad y = 0, 1, 2, \dots, \tag{3.2}$$

where $\mu > 0$. We denote this reparameterized distribution as COM-Poisson$_\mu$. In Figure 2, we show the shapes of the COM-Poisson$_\mu$ distribution.

The constraint $\mu > 0$ imposes an undesirable restriction in the original parameter space, $\lambda^\nu > (\nu - 1)/(2\nu)$. However, this restricted region is related to small underdispersed counts (averages smaller than 0.1 and $\nu > 1$) a parameter region unlikely to be of interest in practice.

In order to explore the flexibility of the COM-Poisson model to deal with real count data,

[2] $\mathrm{E}(Y) \approx \sum_{y=0}^{500} yp(y)$ and $\mathrm{E}(Y^2) \approx \sum_{y=0}^{500} y^2 p(y)$.

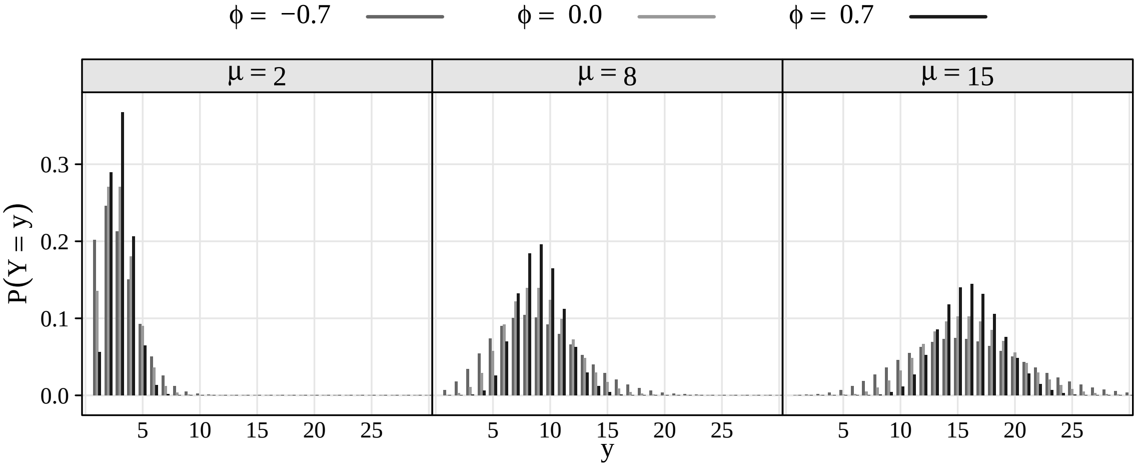

Figure 2: Shapes of the COM-Poisson distribution for different parameter values.

we compute indexes for dispersion (DI), zero-inflation (ZI) and heavy-tail (HI), which are respectively given by

$$\text{DI} = \frac{\text{Var}(Y)}{\text{E}(Y)}, \quad \text{ZI} = 1 + \frac{\log \Pr(Y = 0)}{\text{E}(Y)} \quad \text{and} \quad \text{HT} = \frac{\Pr(Y = y + 1)}{\Pr(Y = y)} \quad \text{for} \quad y \to \infty.$$

These indexes are defined in relation to the Poisson distribution. Thus, the dispersion index indicates overdispersion for DI > 1, underdispersion for DI < 1 and equidispersion for DI $= 1$. The zero-inflation index indicates zero-inflation for ZI > 0, zero-deflation for ZI < 0 and no excess of zeros for ZI $= 0$. Finally, the heavy-tail index indicates a heavy-tail distribution for HT $\to 1$ when $y \to \infty$. These indexes are discussed by Bonat et al. (2017) to study the flexibility of Poisson-Tweedie distribution, and Puig and Valero (2006) to describe count distributions in general.

Regarding the COM-Poisson$_\mu$ distribution, in Figure 3 we present the relationship between (a) mean and variance, (b–c) the dispersion and zero-inflation indexes for different values of μ and ϕ, and (d) the heavy-tail index for $\mu = 25$ and different values of y and ϕ. Figure 3 shows that the indexes are slightly dependent on the expected values and tend to stabilize for large values of μ. Consequently, the mean and variance relationship Figure 3(a) is proportional to the dispersion parameter ϕ. In terms of moments, this leads to a specification indistinguishable from the quasi-Poisson regression model. The dispersion indexes in Figure 3(b) show that the distribution is suitable to deal to dispersed counts, of course. For the parameter values evaluated the largest DI was 4.21 and smallest was 0.168. Figure 3(c) shows the COM-Poisson can handle a limited amount of zero-inflation, in cases of overdispersion ($\phi < 0$). On the other hand, for $\phi > 0$ (underdispersion) this distribution is suitable to deal with zero-deflated counts. Heavy-tail indexes in Figure 3(d) indicate the distribution is in general a light-tailed distribution, i.e. HT $\to 0$ for $y \to \infty$.

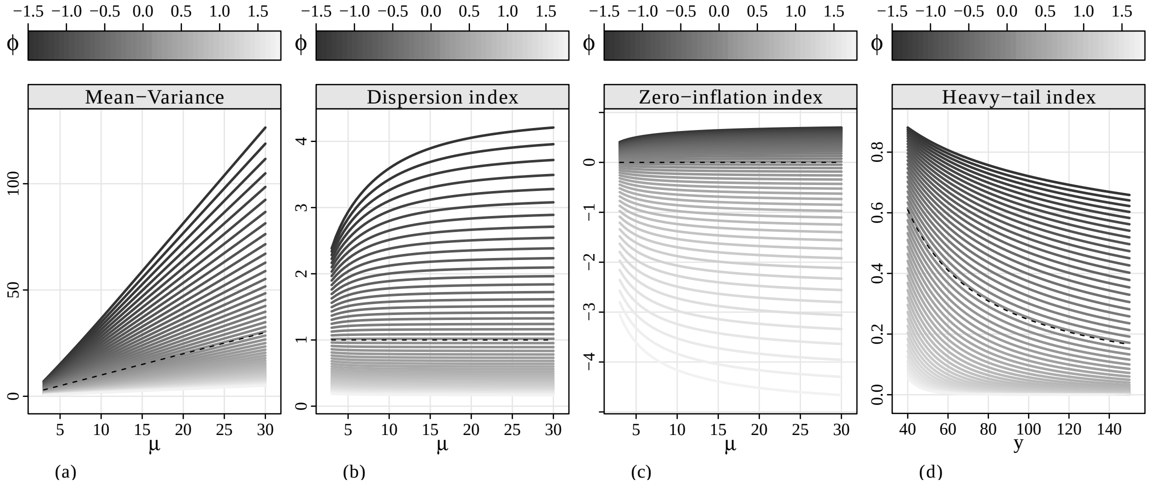

Figure 3: Indexes for the COM-Poisson distribution. (a) Mean and variance relationship, (b–d) dispersion, zero-inflation and heavy-tail indexes for different parameter values. Dotted lines represents the special case of the Poisson distribution.

4 Estimation and Inference

In this section we describe estimation and inference for the two forms of the COM-Poisson regression model based on the maximum likelihood method. Inference can be done using the standard machinery of likelihood inference, including likelihood ratio tests for model comparison and Wald-tests for testing individual (or groups of) parameters. The log-likelihood function for a set of independent observations y_i, $i = 1, 2, \ldots, n$ from the COM-Poisson$_\mu$ distribution has the following form,

$$\ell = \ell(\boldsymbol{\beta}, \phi \mid \boldsymbol{y}) = e^{\phi}\left[\sum_{i=1}^{n} y_i \log\left(\mu_i + \frac{e^{\phi}-1}{2e^{\phi}}\right) - \sum_{i=1}^{n} \log(y_i!)\right] - \sum_{i=1}^{n} \log(Z(\mu_i, \phi)), \tag{4.1}$$

where $\mu_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$, with $\boldsymbol{x}_i^\top = (x_{i1},\, x_{i2},\, \ldots,\, x_{ip})$ is a vector of known covariates for the i-th observation, and $(\boldsymbol{\beta},\, \phi) \in \mathbb{R}^{p+1}$. The normalizing constant $Z(\mu_i, \phi)$ is given by

$$Z(\mu_i, \phi) = \sum_{j=0}^{\infty}\left[\left(\mu_i + \frac{e^{\phi}-1}{2e^{\phi}}\right)^{je^{\phi}} \frac{1}{(j!)^{e^{\phi}}}\right]. \tag{4.2}$$

The evaluation of the log-likelihood function for each observation involves the computation of the infinite series (Equation 4.2). Thus, the fitting procedure is computationally expensive for regions of the parameter space where the convergence of the infinite sum is slow.

Parameter estimation requires the numerical maximization of Equation 4.1. Since the derivatives of ℓ cannot be obtained in closed forms, we use the `BFGS` algorithm (Nocedal and Wright, 1995) as implemented in the function `optim()` in `R` (R CORE TEAM, 2017). Standard errors for the regression coefficients are obtained based on the observed information matrix $\mathcal{I}(\boldsymbol{\theta})$, where $\mathcal{I}(\boldsymbol{\theta}) = -\mathcal{H}(\boldsymbol{\theta})$ (Hessian matrix) is computed numerically by central finite differences. Standard errors for $\hat{\eta}_i = \log(\hat{\mu}_i)$ and hence confidence intervals are obtained by using the delta method (Pawitan, 2001, p. 89).

Parameter estimation for the COM-Poisson regression model in the original parametrization is analogous to that presented for the COM-Poisson$_\mu$ distribution, however, it uses Equation 4.1 re-expressed in terms of λ. Here, even for the standard COM-Poisson distribution, the dispersion parameter is taken on the log scale to avoid numerical issues.

For the applications we also fitted the quasi-Poisson model (Wedderburn, 1974) as a baseline model. This approach is based only on a second-moment assumption and without specific underlying probablity model is less restrictive. In this model the variance of the response variable is specified by an additional dispersion parameter σ, with $\text{Var}(Y_i) = \sigma\mu_i$. These models are fitted in `R` using the function `glm(..., family = quasipoisson)`.

5 Simulation study

In this section we report a simulation study to assess the properties of the maximum likelihood estimators, the approximate parameter orthogonality of the reparametrized model, as well as the flexibility of the COM-Poisson regression model to deal with non-equidispersed count data.

We considered average counts varying from 3 to 27 arising from a regression model with a continuous and a categorical covariate. The continuous covariate ($\boldsymbol{x}_1$) was generated as a linearly increasing sequence from 0 to 1 with length equal to the sample size. Similarly, the categorical covariate ($\boldsymbol{x}_2$) was generated as a sequence of three values each one repeated $n/3$ times (rounding up when required), where n denotes the sample size. The parameter μ of the reparametrized COM-Poisson random variable is given by $\boldsymbol{\mu} = \exp(\beta_0 + \beta_1\boldsymbol{x}_1 + \beta_{21}\boldsymbol{x}_{21} + \beta_{22}\boldsymbol{x}_{22})$, where $\boldsymbol{x}_{21}$ and $\boldsymbol{x}_{22}$ are dummy representing the levels of $\boldsymbol{x}_2$. The regression coefficients were fixed at the values, $\beta_0 = 2$, $\beta_1 = 0.5$, $\beta_{21} = 0.8$ and $\beta_{22} = -0.8$.

We designed four simulation scenarios by considering different values of the dispersion parameter $\phi = -1.6, -1.0, 0.0$ and 1.8. Thus, we have strong and moderate overdispersion, equidispersion, and underdispersion, respectively. Figure 4 shows the variation of the average counts (left) and dispersion index (right) for each value of the dispersion parameter considered in the simulation study. These configurations allow us to assess the properties of the maximum likelihood estimators not only in extreme situations, such as high counts and low dispersion, and low counts and high dispersion, but also in the standard case of equidispersion.

In order to check the consistency of the estimators we considered four different sample sizes: 50, 100, 300 and 1000; generating 1000 data sets in each case. For sample sizes 50 and 100, we have 29 and 8 simulations of the 1000 where the fitting algorithm did not converge. These non-convergence situations occurred for $\phi = -1.6$. In Figure 5, we show the bias of the estimators for each simulation scenario (combination between values of the dispersion

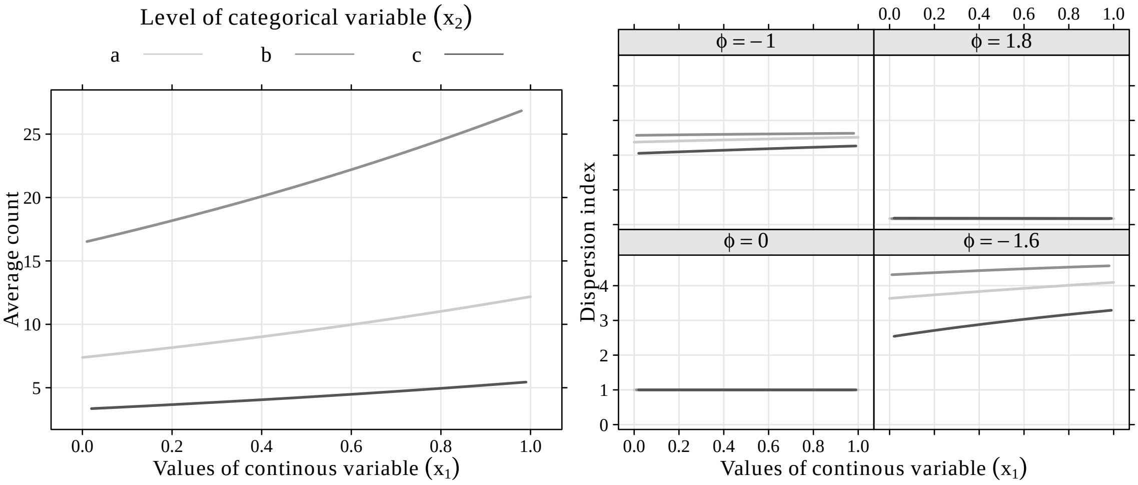

Figure 4: Average counts (left) and dispersion indexes (right) for each scenario considered in the simulation study.

parameter and samples sizes) along with the confidence intervals calculated as average bias plus and minus 1.96 times the average standard error. The scales are standardized for each parameter by dividing the average bias by the average standard error obtained for the sample of size 50.

The results in Figure 5 show that for all dispersion levels, both the average bias and standard errors tend to 0 as the sample size increases. Thus the estimators for the regression parameters are unbiased, consistent and their empirical distributions are symmetric. For the dispersion parameter, the estimator is asymptotically unbiased; in small samples the parameter is overestimated and the empirical distribution is slightly right-skewed.

Figure 6 presents the empirical coverage rate of the asymptotic confidence intervals. The results show that for the regression parameters the empirical coverage rates are close to the nominal level of 95% for sample sizes greater than 100 and all simulation scenarios. For the dispersion parameter the empirical coverage rates are slightly lower than the nominal level; however, they become closer to the nominal level for large samples. The worst scenario is when we have a small sample size and strong overdispersed counts.

To check the orthogonality property we compute the covariance matrix between maximum likelihood estimators $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \phi)$, obtained from the observed information matrix, $\text{Cov}(\hat{\boldsymbol{\theta}}) = \mathcal{I}^{-1}(\boldsymbol{\theta})$. Figure 7 shows the correlation between the regression and dispersion parameter estimators for each simulation scenario, on the correlation scale. The correlations are gnerally close to zero in all cases suggesting the orthogonality property for the reparametrized model. Interestingly, results in the first panel show that $\text{Corr}(\hat{\beta}_{22}, \hat{\phi})$ is not very close to zero (values between -0.4 and 0.2) for strong overdispersion ($\phi = -1.6$).

To illustrate the orthogonality, Figure 8 displays contour plots of the deviance surfaces

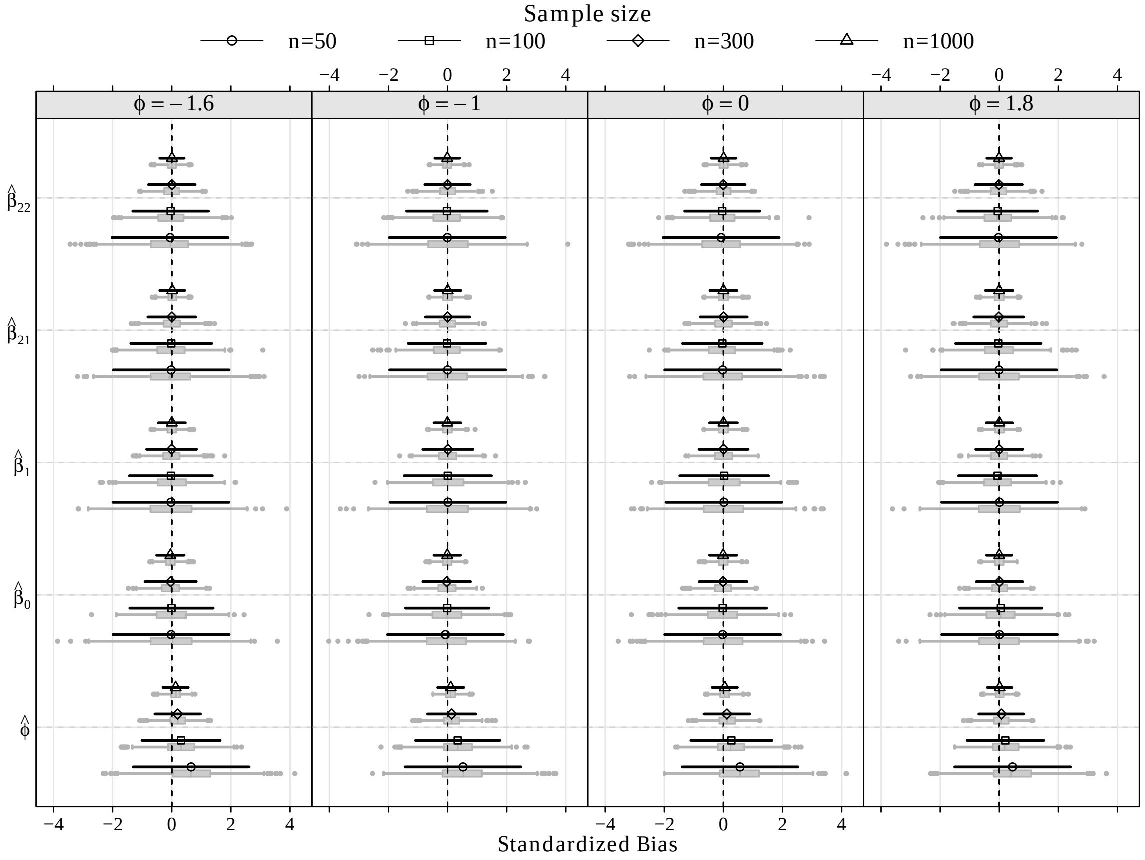

Figure 5: Distributions of standardized bias (gray box-plots) and average with confidence intervals (black segments) by different sample sizes and dispersion levels.

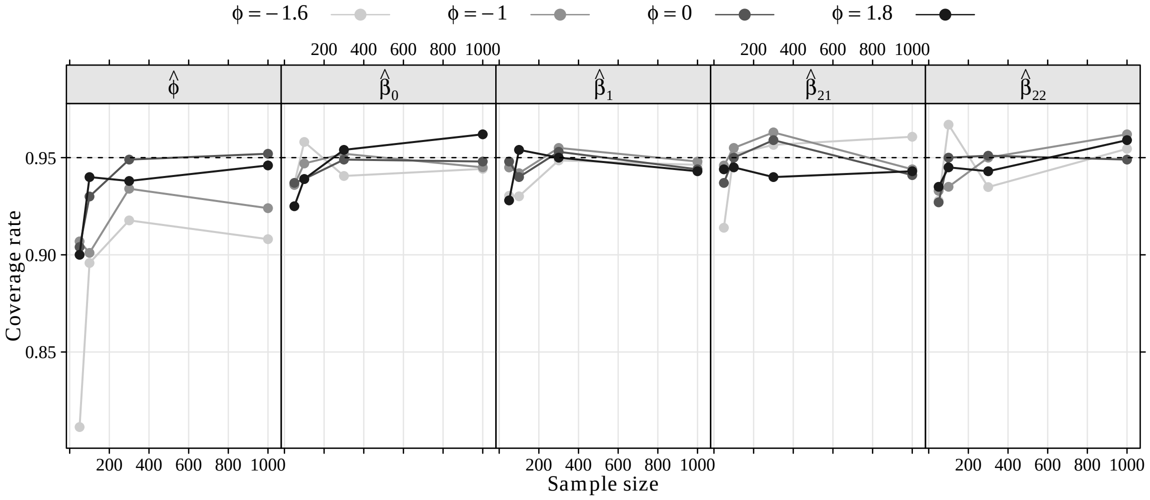

Figure 6: Coverage rate based on confidence intervals obtained by quadratic approximation for different sample sizes and dispersion levels.

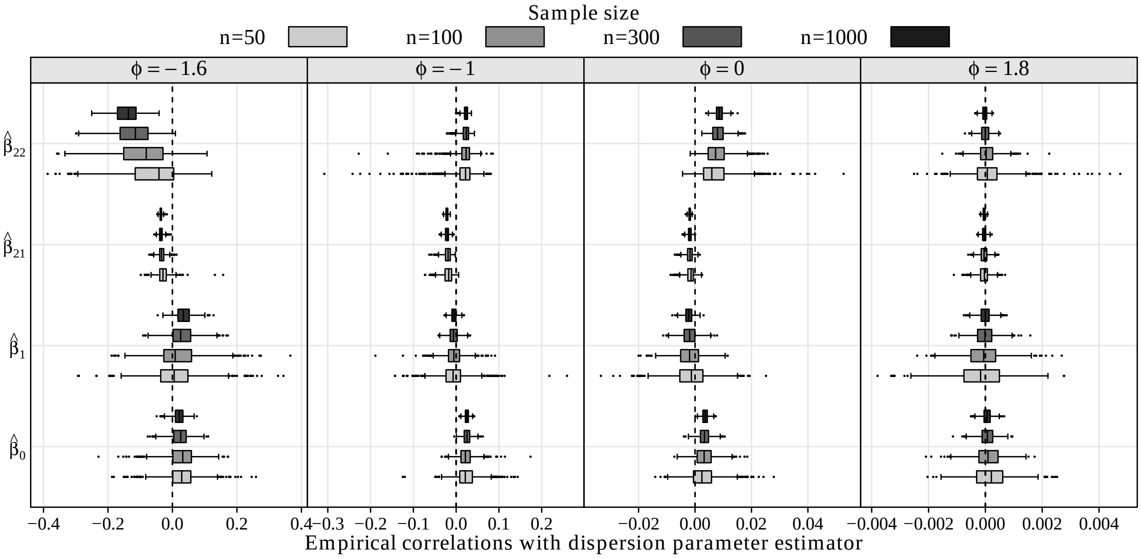

Figure 7: Empirical correlations between regression and dispersion parameters by different sample sizes and dispersion levels.

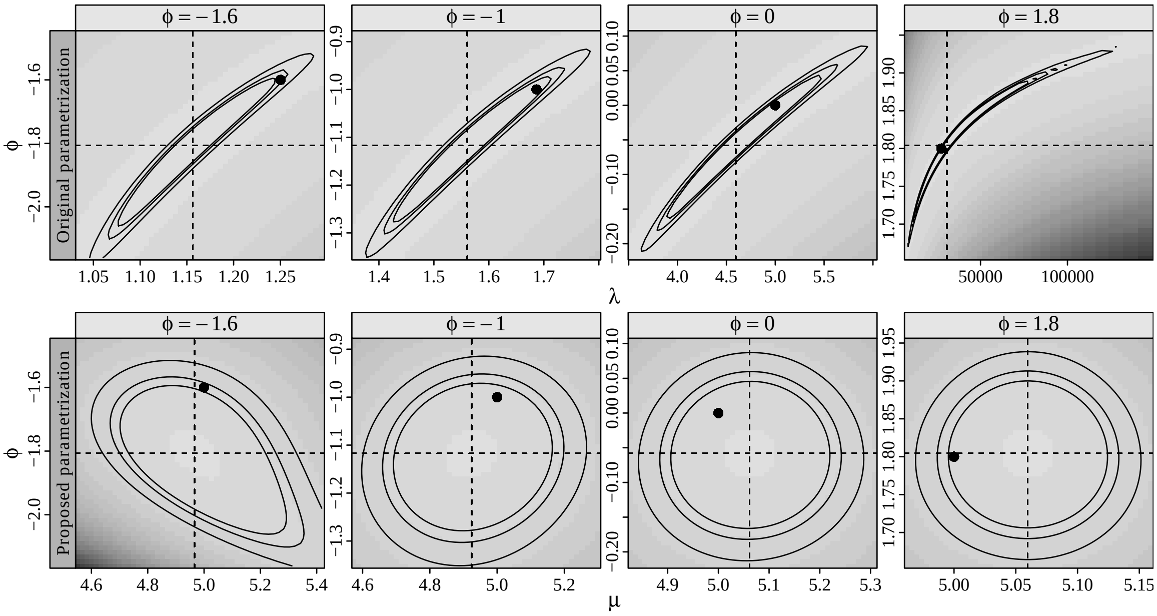

Figure 8: Deviance surfaces contour plots under original and proposed parametrization for four simulated data sets of size 1000 with different dispersion parameters. The ellipses are confidence regions (90, 95 and 99%), dotted lines are the maximum likelihood estimates, and points are the real parameters used in the simulation.

for four simulated data sets of size 1000 with $\mu = 5$ and different values of the dispersion parameters. The shapes of the deviance function show that the proposed parametrization is better for both computation and asymptotic (normal-based) inference. Furthermore, it is interesting to note that the deviance function shape under strong overdispersion ($\phi = -1.6$) is not as well behaved as the others; this is due to the inaccuracy of the mean approximation Figure 1(a). Although the reparametrized model is always valid, μ and ϕ are orthogonal only when μ is the expectation of the distribution, in other words, when the approximation $\lambda - (\nu - 1)/(2\nu)$ is accurate. This also explains the results of $\text{Corr}(\hat{\beta}_{22}, \phi)$ in the first panel of Figure 7, since β_{22} is negative and associated with low counts.

6 Case studies

In this section, we describe three illustrative examples of count data analysis. For the analyses as alternative models we considered the standard Poisson regression model, the COM-Poisson model in the two forms (original and new parametrization), and the quasi-Poisson regression model. The data sets and `R` code for their analysis are available as supplementary material.

6.1 Artificial defoliation in cotton phenology

This example relates to cotton plants (*Gossypium hirsutum*) subjected to five levels of artificial defoliation (`des`) at each of five different growth stages (`est`). The main goal of this study was to assess the effect of defoliation levels at the different growth stages on the cotton production, measured by the number of bolls produced. The experimental study was conducted in a greenhouse and used a completely randomized design with five replicates. This data set was previously analyzed by Zeviani et al. (2014) using the Gamma-Count distribution.

Following Zeviani et al. (2014), the linear predictor is given by

$$\eta_{ij} = \beta_0 + \beta_{1j}\texttt{def}_i + \beta_{2j}\texttt{def}_i^2$$

where $\eta_{ij} = \log(\lambda_{ij})$ for the original COM-Poisson and $\eta_{ij} = \log(\mu_{ij})$ for COM-Poisson$_\mu$. The parameter μ_{ij} is the (approximated) expected number of cotton bolls for the i-th defoliation level ($i = 1$: 0%, 2: 25%, 3: 50%, 4: 75% e 5: 100%) and j-th growth stage ($j =$ 1: vegetative, 2: flower bud, 3: blossom, 4: boll, 5: boll open), i.e. we have a different second order effect of defoliation at each growth stage. The parameter estimates and goodness-of-fit measures for the Poisson, COM-Poisson, COM-Poisson$_\mu$, and quasi-Poisson regression models are presented in Table 2.

The results presented in Table 2 show that the goodness-of-fit measures (log-likelihood, AIC and BIC) are quite similar for the COM-Poisson and COM-Poisson$_\mu$ models. This suggests that the reparametrization does not change the model fit, as is to be expected. The Poisson model is clearly unsuitable, being overly conservative, due to the overestimated standard

Table 2: Parameter estimates (Est) and ratio between estimate and standard error (SE) for the four model strategies for the analysis of the cotton experiment.

	Poisson		COM-Poisson		COM-Poisson$_\mu$		Quasi-Poisson	
	Est	Est/SE	Est	Est/SE	Est	Est/SE	Est	Est/SE
ϕ, σ			1.5847	12.4166	1.5817	12.3922	0.2410	
β_0	2.1896	34.5724	10.8967	7.7594	2.1905	74.6397	2.1896	70.4205
β_{11}	0.4369	0.8473	2.0187	1.7696	0.4350	1.8194	0.4369	1.7260
β_{12}	0.2897	0.5706	1.3431	1.2109	0.2876	1.2227	0.2897	1.1622
β_{13}	-1.2425	-2.0581	-5.7505	-3.8858	-1.2472	-4.4202	-1.2425	-4.1921
β_{14}	0.3649	0.6449	1.5950	1.2975	0.3500	1.3280	0.3649	1.3135
β_{15}	0.0089	0.0178	0.0377	0.0346	0.0076	0.0324	0.0089	0.0362
β_{21}	-0.8052	-1.3790	-3.7245	-2.7754	-0.8033	-2.9613	-0.8052	-2.8089
β_{22}	-0.4879	-0.8613	-2.2647	-1.8051	-0.4858	-1.8499	-0.4879	-1.7544
β_{23}	0.6728	0.9892	3.1347	2.0837	0.6788	2.1349	0.6728	2.0149
β_{24}	-1.3103	-1.9477	-5.8943	-3.6567	-1.2875	-4.0951	-1.3103	-3.9672
β_{25}	-0.0200	-0.0361	-0.0901	-0.0755	-0.0189	-0.0740	-0.0200	-0.0736
LogLik	-255.803		-208.250		-208.398		–	
AIC	533.606		440.500		440.795		–	
BIC	564.718		474.440		474.735		–	

errors. The difference in terms of $-2\times$log-likelihood value from the Poisson to the COM-Poisson$_\mu$ model is 94.811 for one additional parameter, which confirms the significantly improved fit of the COM-Poisson$_\mu$ model. Finally, the estimated value of the dispersion parameter $\hat{\phi} = 1.582$ suggests substantial underdispersion.

Furthermore, results in Table 2 also show the advantage of the COM-Poisson$_\mu$ model, since the regression parameter estimates are quite similar to those obtained for the Poisson model, whereas estimates from the original COM-Poisson model are on a different and not easily interpreted scale. The ratios between estimates and their respective standard errors for the two COM-Poisson models are very close to ratios from the y quasi-Poisson model. However, it is important to note that COM-Poisson model is a full parametric approach, i.e. there is a probability distribution associated to the counts, while the quasi-Poisson model is based only on second-moment assumptions.

Figure 9 presents the observed and fitted values with confidence intervals (95%) as a function of the defoliation level for each growth stage. The fitted values are the same for the Poisson and COM-Poisson$_\mu$ models, however, the confidence intervals are larger for the Poisson model because of the inappropriate equidispersion assumption. The results from the COM-Poisson$_\mu$ model are consistent with those from the Gamma-Count model (Zeviani et al., 2014), Poisson-Tweedie (Bonat et al., 2017) and the alternative parametrization of the COM-Poisson distribution proposed by Huang (2017). All of these models indicated underdispersion and significant effects of defoliation at the vegetative, blossom and boll growth stages.

In order to assess the relation between $\boldsymbol{\mu}$ and ϕ in the COM-Poisson$_\mu$ parametrization, Ta-

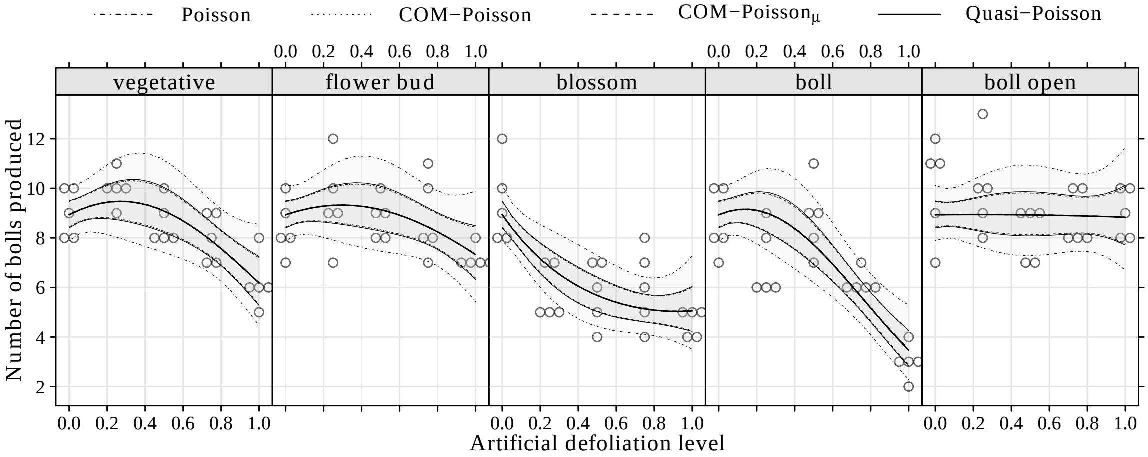

Figure 9: Scatterplots of the observed data and curves of fitted values with 95% confidence intervals as functions of the defoliation level for each growth stage.

ble 3 presents the empirical correlations between the regression and dispersion parameters, as computed by the asymptotic covariance matrix of the estimates, i.e. the inverse of the observed information. The correlations are practically zero for the COM-Poisson$_\mu$, however, with the original parametrization some correlations are quite large, in particular for the parameter β_0 (due to the effects parametrization in the linear predictor). This result explains the improved performance of the maximization algorithm in the new parametrization. It is also important to note that as initial regression parameter values for the `BFGS` algorithm are provided by the Poisson model, then in the COM-Poisson$_\mu$ model the initial values are very close to the maximum likelihood estimates and maximization effort is essentially only on the dispersion parameter ϕ. To compare the computational times for the two parametrizations we repeat the fitting 50 times. In this example the COM-Poisson$_\mu$ fit was, on average, 38% faster than for the original version.

Table 3: Empirical correlations between $\hat{\phi}$ and $\hat{\boldsymbol{\beta}}$ for the two parametrizations of COM-Poisson model fit to underdispersed data.

	$\hat{\beta}_0$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{13}$	$\hat{\beta}_{14}$	$\hat{\beta}_{15}$
COM-Poisson	0.9952	0.2229	0.1526	−0.4895	0.1614	0.0043
COM-Poisson$_\mu$	0.0005	−0.0002	−0.0002	−0.0007	−0.0015	−0.0002
	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$	$\hat{\beta}_{23}$	$\hat{\beta}_{24}$	$\hat{\beta}_{25}$	
COM-Poisson	−0.3496	−0.2276	0.2629	−0.4578	−0.0095	
COM-Poisson$_\mu$	0.0001	0.0002	0.0006	0.0018	0.0001	

6.2 Soil moisture and potassium doses on soybean culture

In this second example we consider a study of potassium doses and soil moisture levels on soybean (*Glicine Max*) production. This was set up as a 5 × 3 factorial experiment in a randomized complete block design with 5 replicates. Five different potassium doses (`K`) (0, 0.3, 0.6, 1.2 and 1.8 × 100mg dm^{-3}) were applied to the soil and soil moisture (`umid`) levels were controlled at (37.5, 50, 62.5%) . The experiment was carried out in a greenhouse and the experimental units were pots with two plants in each. The count response measured was the total number of bean seeds per pot (Serafim et al., 2012). Figure 10 (left) shows the number of bean seeds recorded for each combination of potassium dose and moisture level; it is important to note the indication of a quadratic effect of the potassium levels, as indicated by the smoothers. Most points in the sample variance *versus* sample means dispersion diagram (right) are above the identity line, suggesting overdispersion (block effect not yet removed).

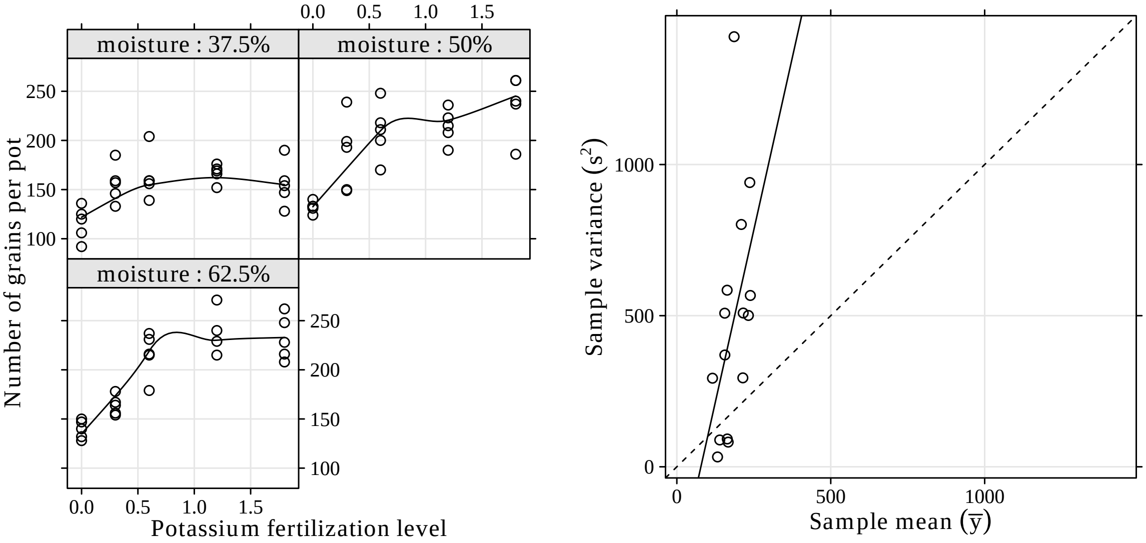

Figure 10: Number of bean seeds per pot for each potassium dose and moisture level (left) and sample mean against sample variance of the five replicates for each experimental treatment (right). Solid lines are smoothing curves (left) and the least-squares line (right).

For the analysis of this data set, based on the descriptive analysis (Figure 10), we proposed the following linear predictor

$$\eta_{ijk} = \beta_0 + \gamma_i + \tau_j + \beta_1 \mathtt{K}_k + \beta_2 \mathtt{K}_k^2 + \beta_{3j} \mathtt{K}_k$$

where γ_i is the effect of i-th block (i =1: block II, 2: block III, 3: block IV and 4: block V), τ_j is the effect of j-th moisture level (j =1: 50% and 2: 62.5%), β_1 and β_2 give the baseline quadratic response over potassium levels `K` (k =1: 0.0, 2: 0.3, 3: 0.6, 4: 1.2, 5: 1.8 100mg dm^{-3}) and β_{3j} is interaction of the first order potassium effect (`K`) for the j-th moisture level (`umid`); The predictor η_{ijk} is $\log(\lambda_{ijk})$ for the original COM-Poisson and $\log(\mu_{ijk})$ for COM-Poisson$_\mu$. Table 4 presents the estimates, ratio between estimate and standard error, and goodness-of-fit measures for the alternative models.

The results in Table 4 show that the two parametrizations of the COM-Poisson model presented very similar goodness-of-fit measures and a better fit than the Poisson model.

The $-2\times$difference between the log-likelihoods of the Poisson and COM-Poisson models was 29.697 on 1 degree of freedom, indicating that ϕ is significantly different from zero. The estimate of ϕ (-0.782) indicates overdispersion, corroborating the descriptive analysis. Concerning the regression parameters, the similarities between the models are analogous to those in the previous section. Both models show effects of block, potassium dose and moisture level, however the Poisson model indicates the effects with greater significance, because it does not take account of the extra variability and so underestimates standard errors.

Table 4: Parameter estimates (Est) and ratio between estimate and standard error (SE) for the four model strategies for the analysis of the soybean experiment.

	Poisson		COM-Poisson		COM-Poisson$_\mu$		Quasi-Poisson	
	Est	Est/SE	Est	Est/SE	Est	Est/SE	Est	Est/SE
ϕ , σ			-0.7785	-4.7208	-0.7821	-4.7371	2.6151	
β_0	4.8666	144.2886	2.2320	6.0415	4.8666	97.7808	4.8666	89.2254
γ_1	-0.0194	-0.7302	-0.0089	-0.4939	-0.0194	-0.4950	-0.0194	-0.4516
γ_2	-0.0366	-1.3733	-0.0169	-0.9212	-0.0366	-0.9306	-0.0366	-0.8492
γ_3	-0.1056	-3.8890	-0.0486	-2.4223	-0.1056	-2.6338	-0.1056	-2.4049
γ_4	-0.0916	-3.2997	-0.0422	-2.1020	-0.0917	-2.2366	-0.0916	-2.0405
τ_1	0.1320	3.6471	0.0609	2.2949	0.1320	2.4715	0.1320	2.2553
τ_2	0.1243	3.4319	0.0573	2.1772	0.1243	2.3258	0.1243	2.1222
β_1	0.6160	11.0139	0.2839	4.7291	0.6161	7.4640	0.6160	6.8108
β_2	-0.2759	-10.2501	-0.1272	-4.5890	-0.2760	-6.9458	-0.2759	-6.3385
β_{31}	0.1456	4.2680	0.0670	2.6138	0.1456	2.8922	0.1456	2.6392
β_{32}	0.1648	4.8294	0.0759	2.8843	0.1648	3.2723	0.1648	2.9864
LogLik	-340.082		-325.241		-325.233		–	
AIC	702.164		674.482		674.467		–	
BIC	727.508		702.130		702.116		–	

The infinite sum $Z(\mu, \phi)$ in the cases of overdispersed count data requires a larger upper bound to reach convergence. Thus, in these cases the computation of the log-likelihood function is computationally expensive. For the data set considered here , the upper bound was fixed at 700. To reach convergence the `BFGS` algorithm evaluated the log-likelihood function 264 times when using the original parametrization of the COM-Poisson distribution and only 20 times with the new parametrization. In terms of computational time, for 50 repetitions of the fit, the proposed reparametrization was, on average, 110% faster than the original one. This is probably due to the better behaviour of the log-likelihood function as well as more relevant initial values from the Poisson fit. In Table 5, we present the empirical correlations between the regression and dispersion parameter estimates. Again, the correlations are close to zero for the COM-Poisson$_\mu$ model, indicating the empirical orthogonality between μ and ϕ.

The observed and fitted counts for each humidity level with confidence intervals are shown in Figure 11. The fitted values are identical for the Poisson and COM-Poisson$_\mu$ models, leading to the same conclusions. On the other hand, confidence intervals for the Poisson

Table 5: Empirical correlations between $\hat{\phi}$ and $\hat{\boldsymbol{\beta}}$ for the two parametrizations of COM-Poisson model fit to overdispersed data.

	$\hat{\beta}_0$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{13}$	$\hat{\beta}_{14}$	$\hat{\beta}_{15}$	$\hat{\beta}_{21}$
COM-Poisson	0.9952	0.2229	0.1526	−0.4895	0.1614	0.0043	−0.3496
COM-Poisson$_\mu$	0.0005	−0.0002	−0.0002	−0.0007	−0.0015	−0.0002	0.0001
	$\hat{\beta}_{22}$	$\hat{\beta}_{23}$	$\hat{\beta}_{24}$	$\hat{\beta}_{25}$			
COM-Poisson	−0.2276	0.2629	−0.4578	−0.0095			
COM-Poisson$_\mu$	0.0002	0.0006	0.0018	0.0001			

model are narrower than those from the COM-Poisson$_\mu$, due to the inappropriate equidispersion assumption of the Poisson model. The confidence intervals from the COM-Poisson$_\mu$ and quasi-Poisson models are also very similar, which again shows the already highlighted similarity between these approaches, however only the COM-Poisson model$_\mu$ corresponds to a fully specified probability model.

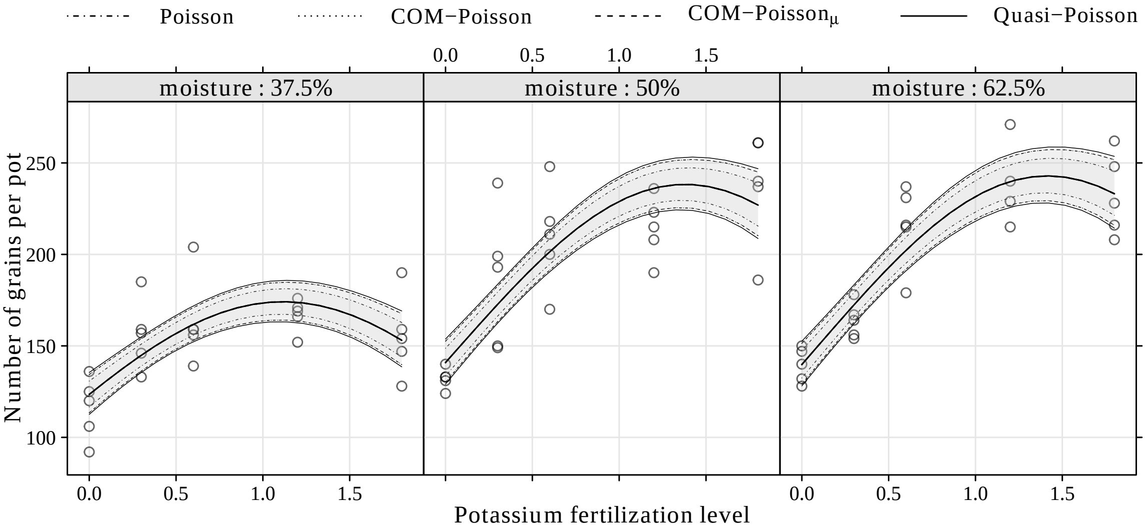

Figure 11: Observed bean seeds counts as functions of potassium doses and soil humidity levels with fitted curves and confidence intervals (95%).

6.3 Assessing toxicity of nitrofen in aquatic systems

Nitrofen is a herbicide that was used extensively for the control of broad-leaved and grass weeds in cereals and rice. Although it is relatively non-toxic to adult mammals, nitrofen is a significant tetragen and mutagen. It is also acutely toxic and reproductively toxic to cladoceran zooplankton. Nitrofen is no longer in commercial use in the U.S., having been the first pesticide to be withdrawn due to tetragenic effects (Bailer and Oris, 1994).

This data set comes from an experiment to measure the reproductive toxicity of the herbicide, nitrofen, on a species of zooplankton (*Ceriodaphnia dubia*). Fifty animals were randomized into batches of ten and each batch was put in a solution with a measured concentration of nitrofen ($0, 0.8, 1.6, 2.35$ and 3.10 μg/10^2litre) (`dose`). Subsequently, the number of live offspring was recorded.

For this data set we consider three models with linear predictors,

$$\begin{aligned}
&\text{Linear:} && \eta_i = \beta_0 + \beta_1\texttt{dose}_i,\\
&\text{Quadratic:} && \eta_i = \beta_0 + \beta_1\texttt{dose}_i + \beta_2\texttt{dose}_i^2 \text{ e}\\
&\text{Cubic:} && \eta_i = \beta_0 + \beta_1\texttt{dose}_i + \beta_2\texttt{dose}_i^2 + \beta_3\texttt{dose}_i^3,
\end{aligned}$$

where $\eta_i = \log(\lambda_i)$ for the original COM-Poisson and $\eta_i = \log(\mu_i)$ for COM-Poisson$_\mu$.

Table 6: Model fit measures and comparisons between predictors and models fitted to the nitrofen data.

Poisson	$\#p$	ℓ	AIC	2(diff ℓ)	diff $\#p$	P($>\chi^2$)	
Preditor 1	2	−180.667	365.335				
Preditor 2	3	−147.008	300.016	67.319	1	2.31E−16	
Preditor 3	4	−144.090	296.180	5.835	1	1.57E−02	
COM-Poisson	$\#p$	ℓ	AIC	2(diff ℓ)	diff $\#p$	P($>\chi^2$)	$\hat{\phi}$
Preditor 1	3	−167.954	341.908				−0.893
Preditor 2	4	−146.964	301.929	41.980	1	9.22E−11	−0.059
Preditor 3	5	−144.064	298.129	5.800	1	1.60E−02	0.048
COM-Poisson$_\mu$	$\#p$	ℓ	AIC	2(diff ℓ)	diff $\#p$	P($>\chi^2$)	$\hat{\phi}$
Preditor 1	3	−167.652	341.305				−0.905
Preditor 2	4	−146.950	301.900	41.405	1	1.24E−10	−0.069
Preditor 3	5	−144.064	298.127	5.773	1	1.63E−02	0.047
Quasi-Poisson	$\#p$	QDev	AIC	F	diff $\#p$	P($>F$)	$\hat{\sigma}$
Preditor 1	3	123.929					2.262
Preditor 2	4	56.610		60.840	1	5.07E−10	1.106
Preditor 3	5	50.774		5.659	1	2.16E−02	1.031

$\#p$, number of parameters; diff ℓ, difference in log-likelihoods; QDev, quasi-deviance, F, F statistics based on quasi-deviances; diff $\#p$, difference in number of parameters.

Table 6 summarizes the results of the fitted models and likelihood ratio tests comparing the sequence of predictors. All models indicate the significance of the cubic effect of the nitrofen concentration. Considering this predictor, there is evidence of equidispersed counts, the ϕ estimate of the COM-Poisson is close to zero and the σ estimate of the quasi-Poisson is close to one. It is interesting to note that if we omit the higher order effects the models show evidence of overdispersion — this exemplifies the discussion on the possible causes of overdispersion in section 1. We also note that the quasi-Poisson approach, although robust to the equidispersion assumption, shows higher descriptive levels (p-values) than parametric models, that is, the tests under parametric models are more powerful than the ones under the quasi-Poisson model in the equidispersed case.

In Table 7, we present the estimates of the regression parameters for the cubic dose models. The interpretations are similar to the other case studies, however, here the Poisson model is also suitable for indicating the significance of the covariate effects. In addition, note that the parameter estimates of the original COM-Poisson model are comparable to the others models. This occurs because we are in the particular case, where $\phi \approx 0$, which implies $\lambda \approx \mu$.

Table 7: Parameter estimates (Est) and ratio between estimate and standard error (SE) for the four model strategies for the analysis of the nitrofen experiment.

	Poisson		COM-Poisson		COM-Poisson$_\mu$		Quasi-Poisson	
	Est	Est/SE	Est	Est/SE	Est	Est/SE	Est	Est/SE
β_0	3.4767	62.8167	3.6494	4.8499	3.4769	64.3083	3.4767	61.8599
β_1	−0.0860	−0.4328	−0.0914	−0.4475	−0.0879	−0.4523	−0.0860	−0.4262
β_2	0.1529	0.8634	0.1612	0.8783	0.1547	0.8938	0.1529	0.8502
β_3	−0.0972	−2.3978	−0.1021	−2.2294	−0.0976	−2.4635	−0.0972	−2.3612

Figure 12 shows the number of live off-spring observed and fitted curves along with confidence intervals for the different cubic dose models. The fitted values and confidence intervals are identical and have a complete overlap. This shows that the estimation of the extra dispersion parameter does not affect the estimation of the regression coefficients in the case of equidispersed counts.

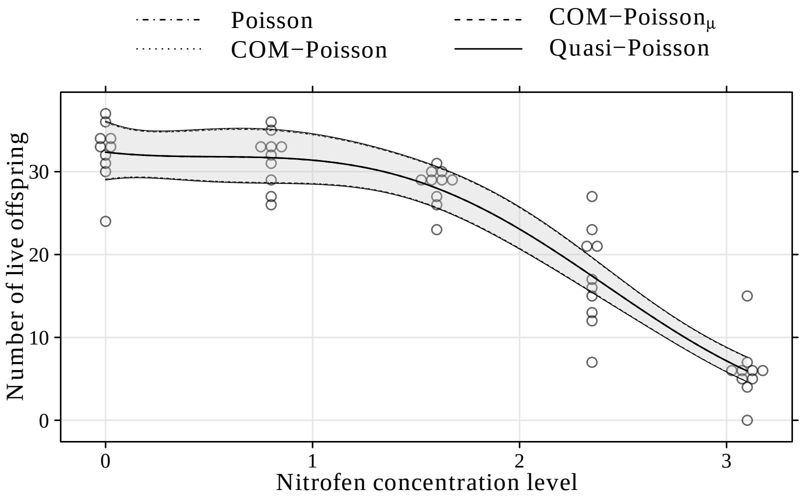

Figure 12: Number of live offspring observed for each nitrofen concentration level with fitted curves and 95% confidence intervals.

Finally, in Table 8 we present the empirical correlations between the regression and dispersion parameter estimates. The results show that even in the special case ($\phi = 0$), the empirical correlations for the original COM-Poisson model are not zero. For the reparametrized model, as discussed in the previous sections, the correlations are practically null. The computational times for fifty repetitions of fit are similar; the average time to fit the COM-Poisson$_\mu$ and COM-Poisson models is 1.19 and 1.09 seconds, respectively.

Table 8: Empirical correlations between $\hat{\phi}$ and $\hat{\boldsymbol{\beta}}$ for the two parametrizations of COM-Poisson model fit to equidispersed data.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
COM-Poisson	0.9972	−0.0771	0.1562	−0.4223
COM-Poisson$_\mu$	−0.0003	0.0023	−0.0029	0.0033

7 Concluding remarks

In this paper, we presented and characterized a novel reparametrization of the COM-Poisson distribution and associated regression model. The reparametrization was based on a simple asymptotic approximation for the expectation of the COM-Poisson distribution. The main advantage of the proposed reparametrization is the simple interpretation of the regression coefficients in terms of the expectation of the response variable as usual in the generalized linear models context. Thus, it is possible to compare directly the results from the COM-Poisson model with those from standard approaches, such as the Poisson and quasi-Poisson regression models. Furthermore, in the new parametrization the COM-Poisson distribution is indexed by μ and an extra dispersion parameter ϕ which our studies suggest are approximately orthogonal. Overall the approach is similar to Huang's (2017) parametrization but ours is simpler, because the μ used here is obtained from simple algebra.

The proposed parametrization is always valid, only the interpretation of μ parameter, as the expectation of the distribution, depends on the accuracy of the approximation. We evaluated the accuracy of the approximations for the expectation and variance of the COM-Poisson distribution by considering quadratic approximation errors. The results showed that the approximation for the expectation is accurate for a large part of the parameter space, supporting our reparametrization. We discuss the properties and flexibility of the distribution to deal with count data through considerations of dispersion, zero-inflation and heavy-tail indexes. A simulation study was used to assess the properties of the reparametrized COM-Poisson model and its ability to deal with different levels of dispersion, as well as the properties of the maximum likelihood estimators. The results of our simulation study suggested that the maximum likelihood estimators of the regression and dispersion parameters are unbiased and consistent. The empirical coverage rates of the confidence intervals computed based on the asymptotic distribution of the maximum likelihood estimators are close to the nominal level for sample sizes greater than 100. The worst case scenario is when we have a small sample size and strongly overdispersed counts. In general, we recommend the use of the asymptotic confidence intervals for computational simplicity, although of course bootstrap intervals could be obtained, but would involve extensive computation.

The three examples and data analyses have shown that the COM-Poisson regression model is a suitable choice to deal with generally dispersed count data, equi-, under- and over-dispersed. The observed empirical correlation between the regression and dispersion parameter estimators and deviance surfaces suggest approximate orthogonality between μ and ϕ in the COM-Poisson$_\mu$ distribution. Thus, the computational procedure based on the proposed reparametrization is faster than that for the original parametrization.

In general, the results presented by the reparametrized COM-Poisson models were satisfactory and comparable to the conventional approaches. Therefore, its use in the analysis of count data is encouraged. The computational routines for fitting the original and reparametrized COM-Poisson regression models are available in the supplementary material[1].

There are many possible extensions to the model discussed in this paper, including simulation studies to assess the model robustness against model misspecification and to assess the theoretical approximations for $Z(\lambda, \nu)$ (or $Z(\mu, \phi)$) (Gaunt et al., 2017). Another simple extension of the proposed model is to model both μ and ϕ parameters as functions of covariates in a double generalized linear model framework. Finally, the reparametrized version of the COM-Poisson model would also be highly suitable for the specification of generalized linear mixed COM-Poisson models.

Acknowledgments

The authors wish to thank the anonymous referees for many comments from which this paper has benefited. This work was partially supported by CNPq, a Brazilian science funding agency.

References

Bailer, A. and Oris, J. (1994). Assessing toxicity of pollutants in aquatic systems. *In Case Studies in Biometry*, pages 25–40.

Bonat, W. H., Jørgensen, B., Kokonendji, C. C., Hinde, J., and Démetrio, C. G. B. (2017). Extended Poisson-Tweedie: properties and regression model for count data. *Statistical Modelling*, **18**(1), 24–49.

Chatla, S. B. and Shmueli, G. (2018). Efficient estimation of COM-Poisson regression and a generalized additive model. *Computational Statistics and Data Analysis*, **121**, 71–89.

Daly, F. and Gaunt, R. E. (2016). The Conway-Maxwell-Poisson distribution: Distributional theory and approximation. *ALEA, Latin American Journal of Probability and Mathematical Statistics*, **13**, 635–658.

Gaunt, R., Iyengar, S., Olde Daalhuis, A., and Simsek, B. (2017). An asymptotic expansion for the normalizing constant of the Conway-Maxwell-Poisson distribution. *Annals of the Institute of Statistical Mathematics*, **to appear**.

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

Huang, A. (2017). Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts. *Statistical Modelling*, **17**(6), 1–22.

Lord, D., Geedipally, S. R., and Guikema, S. D. (2010). Extension of the application of Conway-Maxwell-Poisson models: Analyzing traffic crash data exhibiting underdispersion. *Risk Analysis*, **30**(8), 1268–1276.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, **135**, 370–384.

Nocedal, J. and Wright, S. J. (1995). *Numerical optimization*. Springer. ISBN 0387987932.

Pawitan, Y. (2001). *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press.

Puig, P. and Valero, J. (2006). Count data distributions: some characterizations with applications. *Journal of the American Statistical Association*, **101**(473), 332–340.

R CORE TEAM (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

Serafim, M. E., Ono, F. B., Zeviani, W. M., Novelino, J. O., and Silva, J. V. (2012). Umidade do solo e doses de potássio na cultura da soja. *Revista Ciência Agronômica*, **43** (2), 222–227.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., and Boatwright, P. (2005). A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, **54**(1), 127–142.

Steutel, F. W. and Thiemann, J. G. F. (1989). The gamma process and the Poisson distribution. *(Memorandum COSOR; Vol. 8924). Eindhoven: Technische Universiteit Eindhoven.*

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, **61**(3), 439.

Winkelmann, R. (1995). Duration dependence and dispersion in count-data models. *Journal of Business & Economic Statistics*, **13**(4), 467–474.

Zeviani, W. M., Ribeiro Jr, P. J., Bonat, W. H., Shimakura, S. E., and Muniz, J. A. (2014). The Gamma-count distribution in the analysis of experimental underdispersed data. *Journal of Applied Statistics*, **41**(12), 2616–2626.