Published in final edited form as:

J Appl Stat. 2018 ; 45(5): 799–814. doi:10.1080/02664763.2017.1312299.

Analyzing clustered count data with a cluster specific random effect zero-inflated Conway-Maxwell-Poisson distribution

Hyoyoung Choo-Wosoba^{*,1} and Somnath Datta²

¹Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY 40202, U.S.A.

²Department of Biostatistics, University of Florida, Gainesville, FL 32611, U.S.A.

Abstract

In recent years, data analysis techniques have been developed in biological and medical research areas with different types of count distributions. In particular, zero-inflated versions of parametric count distributions have been used to model excessive zeros that are often present in these assays. Perhaps, the most common count distribution which has been used for analyzing such data is the Poisson distribution. However, a Poisson distribution, having a single underlying parameter, cannot cope with any other data dispersion pattern besides equidispersion. A negative binomial distribution is capable of modeling overdispersed, but not underdispersed data. However, a Conway-Maxwell-Poisson (CMP) distribution (Conway, R. W., and Maxwell, W. L., 1962) can handle not only overdispersion but also underdispersion. We show with an illustrative data set on next generation sequencing of maize hybrids that both underdispersion and overdispersion can be present in genetic data. Furthermore, if count data consists of clustered observations, one of the most efficient statistical technique is to introduce a cluster specific random effect term. Once again, the maize hybrids data presents such a situation.

We develop inference procedures for a zero-inflated CMP regression that incorporates a cluster specific random effect term. Unlike, the Gaussian models, the underlying likelihood is computationally challenging. We use a numerical approximation via a Gaussian quadrature to circumvent this issues. A test for checking zero-inflation has also been developed in our setting. Finite sample properties of our estimators and test have been investigated by extensive simulations. Finally, the statistical methodology has been applied to analyze the maize data mentioned before.

Keywords

Gaussian-Hermite (G-H) quadrature; Mixed effects model; Next- generation sequencing (NGS) data; Poisson distribution; Under- and over-dispersions

*Corresponding author. coolss07@gmail.com.

Conflict of Interest

The authors have declared no conflict of interest.

1 Introduction

Incorporating random effects into a regression model has been a useful statistical tool for the analysis of longitudinal/clustered data. Such random effects allows us to adjust for subject-level randomness. Jointly modeling all sources of variability via mixed effects model typically has more accuracy than a marginal model as long as the random effects structure with random factors is correctly specified. Furthermore, a joint model is able to produce a bona fide likelihood function so that we can carry out inference with standard large sample statistical theory and methods.

Recently, there has been a number of attempts to apply the joint modeling framework to clustered/correlated count data. In many cases, these data include excessive zeros. This is commonly referred to as zero inflation. In general, a Poisson distribution and/or a negative binomial distribution has been applied to a zero-inflated mixed effects model framework (Fulton, Liu, Haynie and Albert, 2015; Hall, 2000; Hasan, Sneddon and Ma, 2009; Yau, Wang and Lee, 2003; Rodrigues-Motta, Gianola and Heringstad, 2010; etc.). However, zero-inflated mixed effects models with these distributions is only suitable for count data with a selective range of data dispersion. As for example, a Poisson distribution, by definition, only fits well for equidispersion, and a negative binomial distribution is only capable of managing overdispersed data.

A Conway-Maxwell-Poisson (CMP) distribution enables us to cope with an unlimited range of dispersion for count data. This distribution includes the commonly used Poisson distribution as a special case and includes only one additional parameter. There has been a recent surge of interest and research activities in the CMP arena, see, e.g., Choo-Wosoba, Levy, and Datta, 2016; Sur, Shmueli, Bose, and Dubey, 2015; Sellers and Raim, 2016; Sellers and Shmueli, 2010; Lord, Guikema, and Geedipally, 2008; Barriga and Louzada, 2014. However, to the best of our knowledge, currently there is no published paper or statistical software programs dealing with a zero-inflated mixed effects analysis with a CMP distribution. Thus, our primary goal in this paper is to explore a mixed effects approach with a zero-inflated Conway-Maxwell-Poisson (ZICMP) distribution. To that end, we adapt the Gaussian-Hermite (G-H) quadrature method to calculate an approximation of the likelihood function from a ZICMP mixed effects model since there is no closed explicit form of the true likelihood function. As a by product of our approach, we are able to construct a statistical test for zero inflation in the data.

The rest of the paper is organized as follows: In Section 2, we provide a brief overview including some important properties of a CMP distribution and describe approximate maximum likelihood estimation of parameters based on the G-H quadrature method mentioned above. Section 3 presents the performance of the point estimators and properties of approximate confidence intervals of the model parameters through an extensive simulation study. We also carry out a numerical power analysis for testing a covariate effect as well as that of a zero-inflation test. In Section 4, we apply our methodology to a maize hybrids experiment dataset to analyze clustered zero-inflated count data with two types of dispersions (over and under). Finally, our paper ends with a conclusion section.

2 Method and Material

A zero-inflated model is composed of two parts: a zero-degenerated distribution and a certain count distribution. The zero-degenerated distribution governs excessive zero values with a Bernoulli distribution and the count distribution, a CMP in this case, governs counts including an expected number of zeros. In particular, the probability mass function (pmf) of a zero-inflated Conway-Maxwell-Poisson (ZICMP) distribution is given by

$$P(Y = y) = \begin{cases} p + (1 - p)/Z(\lambda, \nu), & \text{if } y = 0, \\ (1 - p) \frac{\lambda^y}{(y!)^\nu Z(\lambda, \nu)}, & \text{if } y \geq 1, \end{cases} \quad (1)$$

where λ is a positive rate parameter, ν is a positive dispersion parameter, and

$Z(\lambda, \nu) = \sum_{s=0}^{\infty} \lambda^s / (s!)^\nu$ is a normalizing constant in a regular CMP distribution. It turns out that

for a CMP distribution (Shmueli et al., 2005), ν between 0 and 1 indicates overdispersion of data (i.e., variance is greater than the mean) and $\nu > 1$ indicates underdispersion of data. Furthermore, the ZICMP distribution reduces to a zero-inflated Poisson (ZIP) distribution when $\nu = 1$. In (1), p is the probability that a sample value of the response variable, y , is from a degenerated zero distribution.

We let both the excessive zero probability: and the rate parameter λ depend on covariates through appropriate link functions. The shape parameter ν is held as unknown constant that needs to be estimated from the data as well. We build up correlation in the data by introducing a cluster level random effects in the count data part which is the main component of the distribution for most applications. We feel this modeling offers enough richness for many applications while keeping the likelihood computationally manageable. An extension of this model involving additional random effect terms is discussed in Section 2.4.

In order to keep track of clustered data, we use two different indices, i as a cluster indicator and $j = j(i)$ as an observation indicator within the cluster. The link functions of both parts for a ZICMP joint model is now described as below

$$\log \lambda_{ij} = X_{\beta, ij}^T \beta + \sigma_b b_i, \quad (2)$$

$$\log \frac{p_{ij}}{1 - p_{ij}} = X_{\gamma, ij}^T \gamma,$$

where $X_{\beta, ij}^T \in \mathfrak{R}^{1 \times p_\beta}$ and $X_{\gamma, ij}^T \in \mathfrak{R}^{1 \times p_\gamma}$ are fixed effect covariates corresponding to the j -th observation in the i -th cluster for the count and zero part, respectively. β and γ are fixed effect parameters for their corresponding part. b_i is a random intercept of the i -th cluster and

is normally distributed with mean 0 and standard deviation 1. Therefore, a set of parameters to be estimated in a ZICMP joint model is, then, $\theta = (\beta, \gamma, v, \sigma_b)$.

2.1 Calculation of the approximate likelihood function and estimation of parameters

The likelihood contribution of the i -th cluster in a mixed effect ZICMP model is given by

$$L_i = \int_{-\infty}^{\infty} \left[\prod_{j=1}^{n_i} \left\{ I(Y_{ij} = 0) \left(p_{ij}(\gamma) + \frac{1 - p_{ij}(\gamma)}{Z(\lambda_{ij}(\beta, \sigma_b, b_i), v)} \right) + I(Y_{ij} \geq 1) \left((1 - p_{ij}(\gamma)) \times \frac{\lambda_{ij}(\beta, \sigma_b, b_i)^{y_{ij}}}{(y_{ij}!)^v Z(\lambda_{ij}(\beta, \sigma_b, b_i), v)} \right) \right\} \right] p(b_i) db_i. \quad (3)$$

Note that L_i is calculated not only from multiplying all within-subject (cluster) observations but also by integrating out the subject (cluster) specific random effect. As a result, the observed likelihood function of a cluster in Equation (3) is difficult to compute in a theoretical fashion since there is no closed form of L_i to integrate out with respect to b_i . Thus, we employ a Gaussian-Hermite (G-H) quadrature method to approximate L_i , which is a popular numerical approximation to an integral given by

$$\int_{-\infty}^{\infty} f(x) e^{-x^2} dx \approx \sum_{q=1}^Q f(x_q) w_q, \quad (4)$$

where x_q is a quadrature grid point and w_q is the corresponding weight. We use an R package called fastGHQuad to obtain w_q and x_q . For a direct application of (4) toward calculating L_i , we let, $b_i \sim N(0, 0.5)$ which has $\frac{1}{\sqrt{\pi}} \exp(-b_i^2)$ as its pdf. In addition, the log link function is reparametrized as $\log(\lambda) = X\beta + \sigma_b \sqrt{2} b_i$ which allows us to estimate variance of the random effect term from the count part link function.

Applying this G-H quadrature method in equation (4), an approximation of L_i is given by

$$\tilde{L}_i := \frac{1}{\sqrt{\pi}} \sum_{q=1}^Q w_q \times \left[\prod_{j=1}^{n_i} \left\{ I(Y_{ij} = 0) \left(p_{ij}(\gamma) + \frac{1 - p_{ij}(\gamma)}{Z(\lambda_{ij}(\beta, \sigma_b, x_q), v)} \right) + I(Y_{ij} \geq 1) \left((1 - p_{ij}(\gamma)) \times \frac{\lambda_{ij}(\beta, \sigma_b, x_q)^{y_{ij}}}{(y_{ij}!)^v Z(\lambda_{ij}(\beta, \sigma_b, x_q), v)} \right) \right\} \right], \quad (5)$$

and consequently, an approximate observed log-likelihood function of a mixed effects ZICMP model is obtained as $\log \tilde{L} = \sum_{i=1}^N \log(\tilde{L}_i)$. Parameter estimates are obtained by

maximizing the approximate log-likelihood function. We have used the ‘*optim*’ function in R (version 3.1.2) to this end throughout this paper.

2.2 Variance Estimation

Variance estimation is an integral part of statistical inference based on the approximate MLE obtained above. We have considered both an approximate inverse Fisher information matrix and a sandwich variance estimate. The former is given by

$$\widehat{\text{A-Var}}(\hat{\theta}) = \left(- \sum_{i=1}^N \tilde{H}_i(\hat{\theta}) \right)^{-1},$$

where $\tilde{H}_i(\hat{\theta}) = \frac{\partial^2 \log \tilde{L}_i(\hat{\theta})}{\partial \hat{\theta} \partial \hat{\theta}^T}$ is an approximate Hessian matrix contribution of the i^{th} cluster; it can be directly outputted from the ‘*optim*’ function in R.

The estimated sandwich variance-covariance matrix is calculated using the G-H quadrature approximation is given by

$$\widehat{\text{A-Var}}_S(\hat{\theta}) = \hat{B}^{-1} \hat{M} \hat{B}^{-1},$$

where $\hat{B} = \sum_{i=1}^N \tilde{H}_i(\hat{\theta})$ and $\hat{M} = \sum_{i=1}^N \tilde{S}_i(\hat{\theta}) \tilde{S}_i(\hat{\theta})^T$. Here S_i is obtained as

$$S_i(\theta) = E_{b_i} [S_i(\theta; y_i, b_i) | y_i] = \frac{\int S_i(\theta; y_i, b_i) \psi_i(\theta) f_{b_i}(b_i) db_i}{\int \psi_i(\theta) f_{b_i}(b_i) db_i}$$

where $\psi_i(\theta) = \prod_{j=1}^{n_i} \left\{ I(Y_{ij} = 0) \left(p_{ij}(\gamma) + \frac{1 - p_{ij}(\gamma)}{Z(\lambda_{ij}(\beta, \sigma_b, b_i), v)} \right) + I(Y_{ij} \geq 1) \right\}$, and $f_{b_i}(b_i) = \frac{1}{\sqrt{\pi}} e^{-b_i^2}$ and

$$\times \left(1 - p_{ij}(\gamma) \right) \frac{\lambda_{ij}(\beta, \sigma_b, b_i)^{y_{ij}}}{(y_{ij}!)^v \times Z(\lambda_{ij}(\beta, \sigma_b, b_i), v)} \Bigg\}$$

$\tilde{S}_i$ is its approximation obtained using $\frac{\partial \tilde{L}_i(\theta)}{\partial \theta} \Big|_{\tilde{L}_i(\theta)}$. Note that

$$\begin{aligned} \frac{\partial L_i(\theta)}{\partial \theta} &= \frac{\partial}{\partial \theta} \frac{1}{\sqrt{\pi}} \int \prod_{j=1}^{n_i} \left\{ I(Y_{ij} = 0) \left(p_{ij}(\gamma) + \frac{1 - p_{ij}(\gamma)}{Z(\lambda_{ij}(\beta, \sigma_b, b_i), v)} \right) + \right. & (6) \\ &\left. I(Y_{ij} \geq 1) \left(1 - p_{ij}(\gamma) \right) \frac{\lambda_{ij}(\beta, \sigma_b, b_i)^{y_{ij}}}{(y_{ij}!)^v \times Z(\lambda_{ij}(\beta, \sigma_b, b_i), v)} \right\} \times e^{-b_i^2} db_i. \end{aligned}$$

$$= \frac{1}{\sqrt{\pi}} \int \frac{\partial}{\partial \theta} \psi_i(\theta) e^{-b_i^2} db_i = \frac{1}{\sqrt{\pi}} \int \frac{\partial}{\partial \theta} \log \psi_i(\theta) \times \psi_i(\theta) e^{-b_i^2} db_i \quad (7)$$

$$\approx \frac{1}{\sqrt{\pi}} \sum_{q=1}^Q w_q \frac{\partial}{\partial \theta} \log \psi_i(\theta; x_q) \times \psi_i(\theta; x_q).$$

Thus, $\left. \frac{\partial \tilde{L}_i(\theta)}{\partial \theta} \right|_{\hat{\theta}}$ is finally obtained from Equation (7) (see Appendix A for the details). In

general, it is preferable to use a sandwich variance since it is robust to model misspecification.

2.3 A Statistical Test for Zero-Inflation

Besides testing for a significant regression effect, we may be interested in testing whether there exists zero-inflation in the data. In other words, to determine whether a ZICMP model is statistically necessary over a simpler CMP model under the same mixed effects framework for a given dataset. The parameter associated with zero-inflation is the marginal probability : of excessive zero part. Indeed, it is a testable set of hypotheses given by $H_0 : p = 0$ (no zero inflation) versus $H_1 : p > 0$ (zero-inflation) under the setting where $p \in [0, 1]$ is not covariate dependent.

Note that testing for zero-inflation is different from testing for zero-modification. In fact, likelihood based score tests are applicable and widely used for testing zero-modification $H_1 : p > 0$, but not for zero-inflation. A zero-modified model allows a range of negative values of p so that the value of the null parameter being on the boundary for testing for zero: inflation is avoided (Jansakul and Hinde, 2002). In the zero-inflation test, one of the proper methods for handling the boundary issue is to use likelihood ratio test (LRT) statistics, which asymptotically follows an equal mixture of point mass at 0 and a chi-squared distribution with one degree of freedom (Fulton et. al., 2015). That is,

$\Lambda = 2 \log(L_{\text{ZICMP}}/L_{\text{CMP}}) \xrightarrow{d} 0.5 \times \delta_0 + 0.5 \times \chi_1^2$. However, due to numerical issues in computing the likelihood functions and different fitting procedures of ZICMP and CMP models, one may run into numerical issues in applying the LRT in practice when using the large sample approximation for calibrating the rejection region. Instead, we employ a null bootstrap mechanism to get approximate p-values for the zero-inflation test using the LRT.

The null bootstrap method starts by creating bootstrap (re)samples from the null distribution which asserts the presents of no zero inflation in the data. Thus, while we estimate the model parameters using the original sample, we only resample from the corresponding estimated CMP part obtained by setting p to zero and using the sample estimates for the remaining parameters. For each null bootstrap sample, we can calculate the LRT

$\Lambda_b = 2 \log(L_{b, \text{ZICMP}}/L_{b, \text{CMP}})$ as above using the two maximized log-likelihoods one with ZICMP and the other using CMP, where $b = 1, \dots, B$, where B is a large positive integer, called the bootstrap replication size. Now a bootstrap based p-value is calculated by

$p^* = \sum_{b=1}^B I(\Lambda_{bs} > \Lambda_{obs})/B$, where Λ_{obs} is the value of the LRT for the original sample.

One would reject the null hypothesis if $p^* < \alpha$, α being the nominal level.

2.4 More General Random Effects Models

It is certainly possible to introduce additional random effects into the model corresponding to other sources of variations in both zero and count components. In principle, this will lead to a likelihood contribution for cluster i , involving multiple integration with respect to all the random effects (Min and Agresti, 2005). As before, we could use a Gaussian quadrature for each integral to approximate the likelihood L_i by a sum involving multiple indices. As for example, we could introduce cluster specific random effects in both parts of our ZICMP model by letting

$$\log \lambda_{ij} = X_{\beta,ij}^T \beta + \sigma_b b_i, \quad (8)$$

$$\log \frac{p_{ij}}{1 - p_{ij}} = X_{\gamma,ij}^T \gamma + \sigma_g g_i,$$

where b_i and g_i are independent $N(0, 1)$ variates and σ_b and σ_g are the two unknown variance component parameters. Proceeding as before, we would get an approximate likelihood contribution for the i th cluster as

$$\begin{aligned} \tilde{L}_i := & \frac{1}{\pi} \sum_{q=1}^Q \sum_{q'=1}^{Q'} w_q w_{q'} \times \left[\prod_{j=1}^{n_i} \left(I(Y_{ij} = 0) \left(p_{ij}(\gamma, \sigma_g, x_{q'}) + \frac{1 - p_{ij}(\gamma, \sigma_g, x_{q'})}{Z(\lambda_{ij}(\beta, \sigma_b, x_q), v)} \right) \right. \right. \\ & \left. \left. I(Y_{ij} \geq 1) \left((1 - p_{ij}(\gamma, \sigma_g, x_{q'})) \times \frac{\lambda_{ij}(\beta, \sigma_b, x_q)^{y_{ij}}}{(y_{ij}!)^v \times Z(\lambda_{ij}(\beta, \sigma_b, x_q), v)} \right) \right) \right]. \end{aligned}$$

Parameter estimation and inference can be carried out as before.

3 Simulations

This simulation section contains three types of investigations into the finite sample behaviors of our inferential methodology. First, we study the behaviors of the point estimators in terms of their bias and variance. We also study the performance of approximating normal distribution based confidence intervals for the parameter of interest using two types of variance estimates. Next we conduct a power analysis for a statistical test for a regression effect based on an approximate Wald test; one again both variance estimates have been attempted to standardize the test statistic. The third subsection contains the simulation results for the zero-inflation test. We study the size/power of the test for three choices of the dispersion parameter corresponding to three types of dispersion patterns.

3.1 Study of bias and variance

We considered three different combinations of number of clusters with the same cluster size, namely, $N = 30$, $N = 50$, and $N = 75$ each with $n_i = 10$. Both X_β and X_γ are set as the same design matrix with an intercept term and one continuous covariate randomly generated from a normal distribution with mean 0.94 and variance 0.25, which were arbitrarily chosen. For all three scenarios, we considered the same set of parameter values given by $\beta_0 = 13.8$, $\beta_1 = -1.3$, $\gamma_0 = 2$, $\gamma_1 = -3$, $\nu = 6$, and $\sigma_b^2 = 4$; these values were motivated by an application of CMP regression to airfreight breakage (Kutner, Nachtsheim and Neter, 2003, page 35, Exercise 1.21) by Sellers and Shmueli (2010) although their modeling was less complex than ours since this data set was not clustered and there was no need for considering zero inflation. Given the true parameter values under each scenario, count values including excessive zeros were created. The count values of the i -th cluster from the zero part were randomly sampled from a Bernoulli distribution with success probability, $p_i = \exp(X_{i,\gamma}\gamma) / \exp(X_{i,\gamma}\gamma)$, and the values from the count part were generated by using the inverse CDF transformation method from a CMP distribution with $\lambda_i = \exp(X_{i,\beta}\beta + \sigma_b\sqrt{2}b_i)$ and ν . Here b_i was randomly generated from $N(0, 0.05)$ implying that the true variance of the within cluster random effect was 4 in this simulation.

We have used a Monte-Carlo sample size of $M = 500$ to empirically estimate the bias and the variance of our estimators in each setting. For a single data set, the Gaussian-Hermite (G-H) quadrature method (see Subsection 2.1 for details) was used to approximate L_i via $\tilde{L}_i$. One needs to make a decision about Q , the number of grid points in order to apply Equation 5; for a given Q , both b_{iq} and w_q can be obtained from a R package *fastGHQuad*. Currently, there is no standard statistical criteria to determine an optimal value of Q and a large value of Q may add to the computational burden of the entire process. Following Lesaffre and Spiessens (2001), we choose to use $Q = 25$ in all our subsequent calculations. In fact, an additional analysis (Web Table 1) shows that $Q = 25$ gives the best estimators in terms of bias and SE, compared to other two choices: $Q = 15$ and 35 . The results are given in Table 1, where we report the bias, the true standard error (as obtained from Monte Carlo), and the averages of estimated standard error, using both the information matrix and the sandwich estimator, over all 500 Monte Carlo runs.

As the number of clusters increases, the approximation bias due to the use of a quadrature method seems to dominate the statistical bias since the empirically computed total bias does not converge to zero but rather stays flat. Generally speaking, the empirically estimated true standard deviation (SE) values reduce with sample size. The estimated standard errors for the regression parameters from both methods were reasonably close to the true standard deviations for all scenarios as can be seen from their ratios. However, the sandwich variance method tends to be more conservative than the Fisher information variance method throughout.

Furthermore, we have created probability-probability (p-p) plots to explore the behaviors of the inference based on the asymptotic normality and the two estimates of asymptotic variance in a more precise way. A p-p plot consists of a range of nominal coverage rate

along the x-axis and the corresponding empirical coverage rate along the y-axis and it is a useful tool for checking if a large sample normal approximation of a parameter estimator is effective. The p-p plots from the case of $N = 50$ are shown in Figure 1. For the Fisher information method (the left panel), the p-p plots for most of all the parameters are close to the reference solid line except for σ_b . On the other hand, the p-p plot of σ_b with the sandwich variance method performs noticeably better (the right panel in Figure 1). In addition, the p-p plots corresponding to the remaining parameters behave well in terms of all the points being close to the reference lines. Likewise, for both $N = 30$ and 75 cases, all the p-p plots from the sandwich variance method include points closer to the reference lines than those from the Fisher information method, especially, for σ_b (Web Figures 1 and 2 contain additional supporting information).

3.2 Power analysis for testing a covariate effect

We explore the behavior of statistical power of the Wald test for $H_0 : \beta_1 = 0$, using each of the two variance estimation methods, in all three dispersion cases. Figure 2 describes the power curves corresponding to the two different variance estimation methods. When the true value of β_1 is 0, the empirical size for the sandwich variance method (“sandwich variance” in Figure 2) is 0.054 which is extremely close to the targeted nominal size $\alpha = 0.05$. However, the Fisher information method (“Fisher information” in Figure 2) produces a more inflated size of 0.082. Thus, we can conclude that it is more appropriate to use the sandwich variance method than the inverse of the Fisher information for the approximate maximum likelihood inference in the mixed effects ZICMP regression.

3.3 A power analysis for the zero inflation test

Next, we present the power analysis for the zero-inflation test in three different choices of the dispersion parameters corresponding to underdispersion, overdispersion, and equidispersion for the count part of the model. All the performance of the power analysis is based on $N = 30$ with the same design matrix used for the previous simulation study. Since these analyses are computationally expensive, only two different alternative values of $p_s (= 0.03$ and $0.1)$ are considered for the power of a test along with the null value of $p_s (= 0)$ for each the three dispersion choices.

In this analysis, we use different sets of regression parameters for those three dispersion cases. For the underdispersion case, the set of parameter values is $\{\beta = (5, -2), \nu = 4, \sigma_b = 2\}$. For the overdispersion and equidispersion cases, the sets of values are $\{\beta = (-1.5, 0.5), \nu = 0.6, \sigma_b = 1\}$ and $\{\beta = (0.7, -2), \nu = 1, \sigma_b = 1\}$, respectively. In each case, the parameter values are chosen to maintain a stable range of observations across 500 generated datasets.

Table 4 shows how the power of the ZI test behaves as p increases from its null value of 0 in each of the three different dispersion cases. The empirical size of a test seems to attain the nominal size, $\alpha = 0.05$ for both equidispersion and overdispersion cases. However, the underdispersion case shows the empirical size to be marginally inflated. In addition, compared to p_s , power values increase more dramatically in the underdispersion case compared to the other two cases.

4 Maize hybrids experiment applying next-generation sequencing (NGS) data

This section provides an example to demonstrate the usefulness of a CMP distribution in terms of handling a wide range of dispersion. We use the maize hybrids experiment (Paschold et. al, 2014) recording the expression values of genes in terms of the read counts using a next generation sequencing (NGS) platform. This dataset consists of 39,656 genes with four different hybrid lines of maize (B73, Mo17, B73 $\times$ Mo17, and, Mo17 $\times$ B73). Four experimental units are assigned to each corn type (thus, a total of sixteen corns are used) with four different tissues for each unit. Therefore, there are in total 64 observations for each gene ID. Since these tissues come from four different parts of the same root, the tissues within the same root are possibly correlated leading to clustered count data.

4.1 An underdispersion case

Out of 39,656 genes, the gene ID, “GRMZM2G327208” is selected to represent the case of $\nu > 1$. This gene ID consists of 64 observations including 44 zeros, 9 ones, 8 twos, 1 threes, and 2 fours. A joint model is applied to analyze the data with a ZICMP framework including an random intercept accounting for correlations within each root. Since the total numbers of read counts over genes are different among biological units (in this case, individual replicate for each root), the model needs to be adjusted for normalizing the data by adding an offset term into the link function of the count part. The offset term is calculated by summing all read counts for each biological unit. Thus, our adjusted count part link function becomes $\log(\lambda_{ij}) = X_{ij}^T \beta + \sigma_b \sqrt{2} b_i + \log(\text{offset}_{ij})$, where $i = 1, \dots, 16, j = 1, \dots, 4$. The fixed effect terms on both the count part and zero part are the hybrid types.

A null bootstrap based p-value calculation is used since a normal-based confidence interval may not be appropriate due to a small number (sixteen) of clusters. We generate bootstrapped responses by resampling with replacement of clusters in order to preserve the original correlation within the same roots. Next, we randomly re-assign four different corn types to the sixteen bootstrapped sample roots. One set of bootstrap-based-estimates is obtained after fitting our ZICMP model to a bootstrap sample. This procedure is replicated 500 times. Finally, a bootstrap p-value corresponding to a given hybrid line is given by the proportion these 500 estimates of its regression parameter exceeding the original estimate in absolute value.

The result of a ZICMP joint model analysis on this gene is provided in Table 2. The point estimate of the dispersion parameter, ν , is larger than one, indicating underdispersion from the fixed effects component in the count part. Also, the variance component $\hat{\sigma}_b^2$ is fairly large indicating clustering effects due to different roots. The effect of genotypes does not appear to be statistically significant because all the null bootstrap based p-values are larger than 0.05 in both count and zero parts.

4.2 An over-dispersion case

We select the gene ID, “GRMZM2G041055” to represent the overdispersion case. Table 3 summarizes the results including all the estimates based on a mixed effects ZICMP model with the null bootstrap based p-values for the hybrid effects. According to the results, $\hat{v}$ is (much) smaller than 1 which indicates overdispersion from the fixed effects in the count part and the standard error of the random intercept, $\hat{\sigma}_b$ is also relatively large. Based on the p-values, the hybrid effects are not statistically significant.

We also provide the results based on two other choices for the number of grid points, namely, $Q = 15$ and 35 (see Web Tables 2–3).

5 Conclusion

We have considered analyzing clustered count data based on a mixed effects ZICMP model. However, it is numerically cumbersome to calculate the likelihood function based on a ZICMP mixed effects model due to the absence of an explicit closed form of the likelihood function. While using the G-H quadrature method for the likelihood approximation, we notice the importance of choosing an optimal value for the total number of quadrature points, Q described in subsection 2.1 and simulations in Section 3. Through the simulations, we notice that the approximate likelihood results in biased estimation, while the standard errors get smaller with the number of clusters, as expected. The overall inference from this methodology, specially using the sandwich variance calculation is still adequate for the regression parameters as shown by the p-p plots and the power analysis. The likelihood framework is also suitable for developing a test for zero inflation. However, a bootstrap calibration is recommended instead of the asymptotic distribution to overcome the problem of numerical approximation of the likelihood function. On the data analysis front, we have shown that a ZIP model (that does not have the additional dispersion parameter) may lead to false positive findings.

As alternative way to analyze clustered count data in the ZICMP framework is to perform a marginal regression analysis (Choo-Wosoba, Levy and Datta, 2016). Unlike our mixed effects model in this paper, the marginal model does not specify any random effects because the primary concern of marginal modeling is to get marginal inference over a target population. On the other hand, the mixed effects model focuses on individual’s inference in estimating both fixed effects, β , and the random effect, σ_b . Hence, these two models for zero-inflated clustered data do not support the same statistical inference. In fact, it is not feasible to compare these two sets of estimators statistically.

A limitation of the mixed effects approach is the difficulty of incorporating a large number of random effects term since summing over multiple summands is computationally expensive and approximations by G-H quadrature to the likelihood function lead to computational bias. A Bayesian approach may be a way around in such situations. We will explore this possibility elsewhere.

Finally, we would also like to point out that our model considers a common dispersion parameter over the entire dataset because this setting leading to a parsimonious model is

adequate for most applications. Varying the dispersion parameter with cluster in a general way will lead to an identifiability issue. Certainly, we can let it depend on a covariate (e.g., a group level) through a suitable model if there are compelling reasons to do this. In such cases, one may face additional modeling and computational/convergence issues.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by National Institutes of Health grants 1R03DE020839-01A1 and 1R03DE022538-01. We would like to thank Dan Nettleton for sharing the maize hybrids data and also thank Stephen Cohen for editorial assistance.

Appendix A: Score function of the mixed effect (joint) model based on a ZICMP framework for θ : β , γ , v and σ_b

1.

$$\begin{aligned} \frac{\partial L_i}{\partial \beta} &= \frac{1}{\sqrt{\pi}} \int \left[\frac{\partial}{\partial \beta} \sum_{j=1}^{n_i} I(Y_{ij} = 0) \log \left(p_{ij}(\gamma) + \frac{1 - p_{ij}(\gamma)}{Z(\lambda_{ij}(\beta, \sigma_b, b_{iq}), v)} \right) \right. \\ &\quad \left. + I(Y_{ij} \geq 1) \left\{ \log(1 - p_{ij}(\gamma)) + y_{ij} \log \lambda_{ij}(\beta, \sigma_b, b_{iq}) - \log Z(\lambda_{ij}(\beta, \sigma_b, b_{iq}), v) - v \log(y_{ij}!) \right\} \right] \psi_i \\ &\quad (\theta) e^{-b_i^2} db_i \\ &\approx \frac{1}{\sqrt{\pi}} \sum_{q=1}^Q \left[w_q \left\{ \sum_{j=1}^{n_i} I(Y_{ij} = 0) \frac{(1 - p_{ij}(\gamma)) \sum_{s=0}^{\infty} \frac{s \lambda_{ij}(\beta, \sigma_b, b_{iq})^s}{(s!)^v} X_{ij}}{\left(p_{ij}(\gamma) + \frac{1 - p_{ij}(\gamma)}{Z(\lambda_{ij}(\beta, \sigma_b, b_{iq}), v)} \right) Z(\lambda_{ij}(\beta, \sigma_b, b_{iq}), v)^2} + I(Y_{ij} \geq 1) \right. \right. \\ &\quad \left. \left. \left(y_{ij} - \frac{\sum_{s=0}^{\infty} \frac{s \lambda_{ij}(\beta, \sigma_b, b_{iq})^s}{(s!)^v} \right) \frac{X_{ij}}{Z(\lambda_{ij}(\beta, \sigma_b, b_{iq}), v)} \right\} \psi_i(\theta; b_{iq}) \right] \end{aligned}$$

$$\begin{aligned}
2. \quad \frac{\partial L_i}{\partial \gamma} &= \frac{1}{\sqrt{\pi}} f \left[\sum_{j=1}^{n_i} I(Y_{ij} = 0) \left(\frac{p_{ij}(\gamma)(1 - p_{ij}(\gamma))(1 - Z^{-1}(\lambda_{ij}(\beta, \sigma_b, b_i), v))}{p_{ij}(\gamma) + \frac{1 - p_{ij}(\gamma)}{Z(\lambda_{ij}(\beta, \sigma_b, b_i), v)}} Z_{ij} \right) + I(Y_{ij} \geq 1) \times \right. \\
&\quad \left. - \frac{p_{ij}(\gamma)(1 - p_{ij}(\gamma))}{1 - p_{ij}(\gamma)} Z_{ij} \right] \psi_i(\boldsymbol{\theta}) e^{-b_i^2} db_i \\
&\approx \frac{1}{\sqrt{\pi}} \sum_{q=1}^Q w_q \left[\sum_{j=1}^{n_i} I(Y_{ij} = 0) \left(\frac{p_{ij}(\gamma)(1 - p_{ij}(\gamma))(1 - Z^{-1}(\lambda_{ij}(\beta, \sigma_b, b_{iq}), v))}{p_{ij}(\gamma) + \frac{1 - p_{ij}(\gamma)}{Z(\lambda_{ij}(\beta, \sigma_b, b_{iq}), v)}} Z_{ij} \right) \right. \\
&\quad \left. + I(Y_{ij} \geq 1) \times - \frac{p_{ij}(\gamma)(1 - p_{ij}(\gamma))}{1 - p_{ij}(\gamma)} Z_{ij} \right] \psi_i(\boldsymbol{\theta}; b_{iq})
\end{aligned}$$

$$\begin{aligned}
3. \quad \frac{\partial L_i}{\partial v} &= \frac{1}{\sqrt{\pi}} f \left[\sum_{j=1}^{n_i} I(Y_{ij} = 0) \frac{-(1 - p_{ij}(\gamma)) \sum_{s=0}^{\infty} \frac{\log(s!) \lambda_{ij}(\beta, \sigma_b, b_i)^s}{(s!)^v}}{-\left(p_{ij}(\gamma) + \frac{1 - p_{ij}(\gamma)}{Z(\lambda_{ij}(\beta, \sigma_b, b_i), v)}\right) Z(\lambda_{ij}(\beta, \sigma_b, b_i), v)^2} + I(Y_{ij} \geq 1) \right. \\
&\quad \left. \times \left(\frac{\sum_{s=0}^{\infty} \frac{\log(s!) \lambda_{ij}(\beta, \sigma_b, b_i)^s}{(s!)^v}}{Z(\lambda_{ij}(\beta, \sigma_b, b_i), v)} - \log(y_{ij}!) \right) \right] \psi_i(\boldsymbol{\theta}) e^{-b_i^2} db_i \\
&\approx \frac{1}{\sqrt{\pi}} \sum_{q=1}^Q w_q \left[\sum_{j=1}^{n_i} I(Y_{ij} = 0) \frac{-(1 - p_{ij}(\gamma)) \sum_{s=0}^{\infty} \frac{\log(s!) \lambda_{ij}(\beta, \sigma_b, b_{iq})^s}{(s!)^v}}{-\left(p_{ij}(\gamma) + \frac{1 - p_{ij}(\gamma)}{Z(\lambda_{ij}(\beta, \sigma_b, b_{iq}), v)}\right) Z(\lambda_{ij}(\beta, \sigma_b, b_{iq}), v)^2} \right. \\
&\quad \left. + I(Y_{ij} \geq 1) \times \left(\frac{\sum_{s=0}^{\infty} \frac{\log(s!) \lambda_{ij}(\beta, \sigma_b, b_{iq})^s}{(s!)^v}}{Z(\lambda_{ij}(\beta, \sigma_b, b_{iq}), v)} - \log(y_{ij}!) \right) \right] \psi_i(\boldsymbol{\theta}; b_{iq})
\end{aligned}$$

$$\begin{aligned}
4. \quad \frac{\partial L_i}{\partial \sigma_b} &= \frac{1}{\sqrt{\pi}} \int \left\{ \sum_{j=1}^{n_i} I(Y_{ij} = 0) \frac{(1 - p_{ij}(\gamma)) \sqrt{2} b_i \sum_{s=0}^{\infty} \frac{s \lambda_{ij}(\beta, \sigma_b, b_i)^s}{(s!)^v}}{-\left(p_{ij}(\gamma) + \frac{1 - p_{ij}(\gamma)}{Z(\lambda_{ij}(\beta, \sigma_b, b_i), v)} \right) Z(\lambda_{ij}(\beta, \sigma_b, b_i), v)^2} + I(Y_{ij} \geq 1) \right. \\
&\quad \times \left. \sqrt{2} b_i \left(y_{ij} - \frac{\sum_{s=0}^{\infty} \frac{s \lambda_{ij}(\beta, \sigma_b, b_i)^s}{(s!)^v}}{Z(\lambda_{ij}(\beta, \sigma_b, b_i), v)} \right) \right\} \psi_i(\theta) e^{-b_i^2} db_i \\
&\approx \frac{1}{\sqrt{\pi}} \sum_{q=1}^Q w_q \left\{ \sum_{j=1}^{n_i} I(Y_{ij} = 0) \frac{(1 - p_{ij}(\gamma)) \sqrt{2} b_{iq} \sum_{s=0}^{\infty} \frac{s \lambda_{ij}(\beta, \sigma_b, b_{iq})^s}{(s!)^v}}{-\left(p_{ij}(\gamma) + \frac{1 - p_{ij}(\gamma)}{Z(\lambda_{ij}(\beta, \sigma_b, b_{iq}), v)} \right) Z(\lambda_{ij}(\beta, \sigma_b, b_{iq}), v)^2} \right. \\
&\quad \left. + I(Y_{ij} \geq 1) \times \sqrt{2} b_{iq} \left(y_{ij} - \frac{\sum_{s=0}^{\infty} \frac{s \lambda_{ij}(\beta, \sigma_b, b_{iq})^s}{(s!)^v}}{Z(\lambda_{ij}(\beta, \sigma_b, b_{iq}), v)} \right) \right\} \psi_i(\theta; b_{iq})
\end{aligned}$$

REFERENCES

- Blocker, A.NW., 2014 fastGHQuad: Fast Rcpp implementation of Gauss-Hermite quadrature. <https://cran.rproject.org/web/packages/fastGHQuad/fastGHQuad.pdf>
- Choo-Wosoba H, Levy SM, and Datta S, 2016 Marginal regression models for clustered count data based on zero-inflated Conway-Maxwell-Poisson distribution with applications. *Biometrics* 72, 606–618. [PubMed: 26575079]
- Conway RW, and Maxwell WL, 1962 A queuing model with state dependent service rates. *Journal of Industrial Engineering* 12, 132–136.
- Fulton KA, Liu D, Haynie DL, and Albert PS, 2015 Mixed model and estimating equation approaches for zero inflation in clustered binary response data with application to a dating violence study. *The Annals of Applied Statistics* 9, 275–299. [PubMed: 26937263]
- Barriga Gladys D.C. and Francisco Louzada, 2014 The zero-inflated Conway-Maxwell-Poisson distribution: Bayesian inference, regression modeling and influence diagnostic. *Statistical Methodology* 21, 23–34.
- Hall DB, 2000 Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics* 56, 1030–1039. [PubMed: 11129458]
- Hasan MT, Sneddon G, and Ma R, 2009 Pattern-mixture zero-inflated mixed models for longitudinal unbalanced count data with excessive zeros. *Biometrics* 51, 946–960.
- Jansakul N, and Hinde JP, 2002 Score tests for zero-inflated Poisson models. *Computational Statistics & Data Analysis* 40, 75–96.
- Kutner MH, Nachtsheim CJ, and Neter J, 2003 *Applied Linear Regression Models*, 4th ed., McGraw-Hill, New York.
- Lesaffre E, and Spiessens B, 2001 On the effect of the number of quadrature points in a logistic random-effects model: an example. *Applied Statistics* 50, 325–335.
- Lord D, Guikema SD, and Geedipally SR, 2008 Application of the Conway-Maxwell-Poisson generalized linear model for analyzing motor vehicle crashes. *Accident Analysis & Prevention* 40, 1123–1134. [PubMed: 18460381]
- Min Y, and Agresti A, 2005 Random effect models for repeated measures of zero-inflated count data, *Statistical Modelling*, 5, 1–19.

- Paschold A, Larson NB, Marcon C, Schnable JC, Yeh CT, Lanz C, Nettleton D, Piepho HP, Schnable PS, and Hochholdinger F, 2014 Nonsyntenic genes drive highly dynamic complementation of gene expression in maize hybrids. *The Plant Cell* 26, 3939–3948. [PubMed: 25315323]
- Rodrigue-Motta M, Gianola D, and Heringstad B, 2010 A mixed effects model for overdispersed zero inflated Poisson data with an application in animal breeding. *Journal of Data Science* 8, 379–396.
- Sellers KF, and Raim A, 2016 A flexible zero-inflated model to address data dispersion. *Computational Statistics and Data Analysis* 99, 68–80.
- Sellers KF, and Shmueli G, 2010 A flexible regression model for count data, *The Annals of Applied Statistics* 4, 943–961.
- Shmueli G, Minka TP, Kadane JB, Borle S, and Boatwright P, 2005 A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution. *Annals of Applied Statistics* 54, 127–142.
- Sur P, Shmueli G, Bose S, and Dubey Paromita, 2015 Modeling bimodal discrete data using Conway-Maxwell-Poisson mixture models, *Journal of Business and Economic Statistics*, 33, 352–365.
- Yau KKW, Wang K, and Lee AH, 2003 Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrics* 45, 437–452.

FIGURE 1.

The p-p plot of confidence intervals of all parameters based on our simulation models when the number of clusters is 30. The left panel is based on the inverse of the Fisher information matrix and the right panel is based on the sandwich variance estimator.

FIGURE 2.

The power plots for testing the effect of X_1 using two types of asymptotic variance estimation methods. The red curve corresponds to using the inverse of the Fisher information and the blue curve corresponds to using a sandwich variance estimator. The horizontal black line denotes the nominal size.

Table 1.

Finite sample behavior of parameter estimators obtained from fitting ZICMP mixed effect model when the likelihood is approximated by G-H quadrature with 25 grid points. Here SE_{Fisher} indicates estimated standard errors based on the inverse of the Fisher information matrix and SE_{sand} refers to estimated standard errors from a sandwich covariance variance matrix. Also, SE stands for the standard error based on the Monte Carlo replicates and is taken to be the gold standard for comparison.

		<i>N</i> = 30						<i>N</i> = 50						<i>N</i> = 75					
	True	bias	SE	SE _{Fisher}	SE _{sand}	SE/SE _{Fisher}	SE/SE _{sand}	bias	SE	SE _{Fisher}	SE _{sand}	SE/SE _{Fisher}	SE/SE _{sand}	bias	SE	SE _{Fisher}	SE _{sand}	SE/SE _{Fisher}	SE/SE _{sand}
β_0	13.8	−0.296	1.904	1.500	2.125	1.270	0.896	−0.474	1.531	1.153	1.676	1.327	0.913	−0.537	1.302	0.936	1.324	1.392	0.984
β_1	−1.3	0.014	0.230	0.207	0.272	1.113	0.848	0.037	0.193	0.158	0.209	1.216	0.921	0.046	0.153	0.129	0.167	1.190	0.915
γ_0	2	0.081	0.369	0.355	0.349	1.040	1.058	0.003	0.270	0.269	0.265	1.003	1.019	0.048	0.267	0.222	0.222	1.202	1.202
γ_1	−3	−0.099	0.432	0.400	0.396	1.080	1.091	−0.002	0.304	0.302	0.297	1.005	1.022	−0.053	0.304	0.250	0.250	1.217	1.218
ν	6	−0.060	0.743	0.649	0.935	1.146	0.795	−0.159	0.624	0.495	0.705	1.259	0.885	−0.196	0.531	0.401	0.561	1.323	0.947
σ_b	2	−0.134	0.340	0.227	0.450	1.496	0.757	−0.183	0.276	0.172	0.351	1.601	0.786	−0.217	0.223	0.140	0.311	1.592	0.716

Table 2.

Results for the *GRMZM2G327208* gene from the maize hybrids data, for which the estimated dispersion parameter was larger than 1 representing underdispersion in the fixed parameter count data part. Parameter estimates are reported with corresponding null bootstrap based p-values. Here, the bootstrap replication size was 500.

Mixed effects ZICMP				
	Count part	P-value	Zero part	P-value
<i>Intercept</i>	-17.898	0.044	-6.893	0.178
<i>B73 × Mo17</i>	2.783	0.294	6.794	0.294
<i>Mo17</i>	4.321	0.144	7.286	0.252
<i>Mo17 × B73</i>	3.640	0.178	6.225	0.282
<i>v</i>	2.516			
<i>σ_b</i>	1.156			

Table 3.

Results for the *GRMZM2G041055* gene from the maize hybrids data, for which the estimated dispersion parameter was smaller than 1 representing overdispersion in the fixed parameter count data part. Parameter estimates are reported along with corresponding null bootstrap based p-values. Here, the bootstrap replication size was 500.

Mixed effects ZICMP				
	Count part	P-value	Zero part	P-value
<i>Intercept</i>	-14.887	0.494	-1.795	0.588
<i>B73 × Mo17</i>	-0.574	0.600	-1.217	0.610
<i>Mo17</i>	-2.093	0.132	-1.246	0.562
<i>Mo17 × B73</i>	-0.283	0.748	0.617	0.744
ν	0.601			
σ_b	0.272			

Table 4.

Power of a zero inflation test (nominal size is 5%) based on a ZICMP model for 30 clusters. Three different choices of the dispersion parameter ν was used. Each entry is based on a Monte Carlo sample size of 500.

Proportion of zero inflation	Underdispersion	Choice of ν	Equidispersion	Overdispersion
0	0.092	0.042		0.040
0.03	0.872	0.216		0.120
0.1	1.000	0.476		0.280