DATA DISPERSION: NOW YOU SEE IT... NOW YOU DON'T

Kimberly F. Sellers
Department of Mathematics and Statistics
Georgetown University
Washington, DC 20057
kfs7@georgetown.edu

Galit Shmueli
Indian School of Business
Gachibowli, Hyderabad 500 032
India
galit_shmueli@isb.edu

Key Words: Conway-Maxwell-Poisson (COM-Poisson) regression; mixture model; apparent dispersion; over-dispersion; under-dispersion

ABSTRACT

Poisson regression is the most well-known method for modeling count data. When data display over-dispersion, thereby violating the underlying equi-dispersion assumption of Poisson regression, the common solution is to use negative-binomial regression. We show, however, that count data that appear to be equi- or over-dispersed may actually stem from a mixture of populations with different dispersion levels. To detect and model such a mixture, we introduce a generalization of the Conway-Maxwell-Poisson (COM-Poisson) regression model that allows for group-level dispersion. We illustrate mixed dispersion effects and the proposed methodology via semi-authentic data.

1. INTRODUCTION

Poisson regression is a popular tool for modeling count data, but its underlying assumption of equi-dispersion (i.e., an equal mean and variance) limits its use in many real-world applications with over- or under-dispersed data. Overdispersion frequently arises for various reasons, including mechanisms that generate excessive zero counts or censoring. As a result, over-dispersed count data are common in many areas which, in turn, has lead to the development of statistical methodology for modeling over-dispersed data. Among these models, the negative binomial regression is a popular choice for modeling the relationship between explanatory variables and an outcome count variable. The popularity of the negative binomial regression can be seen by its availability in many statistical software packages and the large literature reporting use of this model.

While over-dispersion is common in count data, under-dispersion also occurs with some frequency. Negative binomial regression, however, is not equipped to model under-dispersed data. A few models exist that allow for both over- and under-dispersion. One example is the restricted generalized Poisson regression of Famoye (1993), which is based on the distribution,

$$P(Y_i = y_i \mid \mu_i, \alpha) = \left(\frac{\mu_i}{1+\alpha\mu_i}\right)^{y_i} \frac{(1+\alpha y_i)^{y_i-1}}{y_i!} \exp\left(\frac{-\mu_i(1+\alpha y_i)}{1+\alpha\mu_i}\right), \; y_i = 0, 1, 2, \ldots, \quad (1)$$

where $\alpha = 0$ represents the special case of a Poisson distribution; $\alpha > 0$ and $-2/\mu_i < \alpha < 0$ represent over- and under-dispersion, respectively. This parameter, however, is notably restricted in the amount of underdispersion that it can model; see Famoye (1993) for discussion. Another model that allows for a range of under- and over-dispersion is the Conway-Maxwell-Poisson (COM-Poisson) regression of Sellers and Shmueli (2010), which generalizes Poisson and logistic regression, and can successfully model count data with a wide range of dispersion levels. It is based on the COM-Poisson distribution, whose probability mass function is

$$P(Y_i = y_i) = \frac{\lambda_i^{y_i}}{(y_i!)^\nu Z(\lambda_i, \nu)}, \quad y_i = 0, 1, 2, \ldots, \quad (2)$$

where $Z(\lambda_i, \nu) = \sum_{s=0}^{\infty} \frac{\lambda_i^s}{(s!)^\nu}$ is the normalizing constant, $\lambda_i = E(Y_i^\nu)$, and $\nu \geq 0$ is the dispersion parameter. The COM-Poisson distribution includes three well-known distributions

as special cases: Poisson ($\nu = 1$), geometric ($\nu = 0, \lambda_i < 1$), and Bernoulli (with probability $\frac{\lambda_i}{1+\lambda_i}$, as $\nu \to \infty$). See Shmueli et al. (2005) for details regarding this distribution. The COM-Poisson distribution also has special properties as it is a weighted Poisson distribution (see Castillo and Perez-Casany, 1998 and 2005; and Kokonendji et al., 2008). Both the COM-Poisson regression and restricted generalized Poisson regression assume a fixed dispersion level, which does not vary across observations (denoted by ν or α, respectively).

In general, two-parameter count data regression models that accommodate over- and/or under-dispersion are based on an underlying distribution whose mean and variance are each a function of both parameters. In other words, the two parameters do not map one-to-one to the mean and variance. This is in stark contrast to the standard linear regression model for continuous data, where the two parameters (μ and σ) correspond to the normal mean and standard deviation that are functionally independent. Whereas mixing normal data with the same μ but different σ levels leads to normally distributed data with the mean μ and a variance that is a "compromise" between the individual variances, the result of mixing count data with multiple dispersion levels leads to data that can appear to have any sort of dispersion.

Given a dataset, how can one determine whether its observed dispersion (or lack thereof) is "real" or the result of a mixture of dispersion levels? In other words, is the observed dispersion level true or "apparent"? Answering this question is important in many applications, and hence there exists extensive statistical literature on mixture models. While mixture models are sometimes used simply for mathematical flexibility, they are also used for modeling a population that is a mixture of sub-populations. We focus here on the latter case, although our proposed model can easily be used for the first purpose as well.

Hilbe (2007) addresses the specific issue of apparent over-dispersion, specifically where equi-dispersed count data appear to be over-dispersed. Noted causes of such apparent over-dispersion include omitting important explanatory variables, outliers, model formulation that does not include sufficient interaction terms, misrepresentation of an explanatory variable in the appropriate scale, and misspecification of the link function (Hilbe, 2007). Correcting for

such causes can remove the apparent over-dispersion, thereby producing an equi-dispersed dataset that can then be fit via Poisson regression. In this paper, we introduce a new important cause that can lead to apparent dispersion of any form (including equi-dispersion), namely mixtures of dispersion levels.

In the context of mixing data with different dispersion levels, Park and Lord (2009) use mixtures of negative binomial regression models with different dispersion parameters (as well as mixtures of Poisson regression models with different mean values) to investigate the nature of over-dispersion found in transportation crash data, and discover that their dataset "seemed to be generated from two distinct sub-populations", each with different levels of over-dispersion. Via a Bayesian formulation and model estimation using MCMC, Park and Lord (2009) encountered several challenges that lead them to the conclusion that "developing a COM-Poisson mixture model may prove to be useful for analyzing motor vehicle crashes." Meanwhile, in marketing research, an important goal is capturing heterogeneity in consumer behavior. Count data arise in many marketing studies, including satisfaction surveys or data on shopping frequency and quantity. Oftentimes, the data arise from different groups of consumers, which can be an important source of heterogeneity; see, e.g., Borle et al. (2007). In such cases, different groups might have data with different levels of dispersion that, if ignored, would lead to a lesser understanding of the heterogeneity sources.

In this paper, we propose a method for detecting apparent dispersion based on COM-Poisson regression. Our approach is novel in two ways. First, we look at mixing not only equi-dispersed and over-dispersed data, but under-dispersed data (or a combination of these). Second, by using the COM-Poisson distribution that belongs to the exponential family in both its parameters, we obtain elegant estimation and inference methods, and avoid the estimation challenges encountered in Park and Lord's Bayesian approach. Our solution is achieved by proposing a generalization of the COM-Poisson regression formulation by Sellers and Shmueli (2010) that allows for group-level dispersion. The model is then used to test for different levels of dispersion across different groups of observations, and to estimate the associated group-level dispersion levels.

The paper is organized as follows. Section 2 briefly describes the COM-Poisson regression model, which assumes a fixed dispersion level across all observations. We then extend this model to account for group-level dispersion such that different groups of observations can have different dispersion levels. Section 3 presents two semi-authentic datasets on elephant matings, where (in each case) a sample of over-dispersed data is combined with a sample of under-dispersed data. We then apply Poisson regression, ordinary COM-Poisson regression, and our proposed group-level dispersion COM-Poisson model to the data to illustrate the results of assuming a particular dispersion structure. Section 4 presents concluding remarks.

2. COM-POISSON REGRESSION

2.1. COM-POISSON REGRESSION WITH FIXED DISPERSION LEVEL

Taking a GLM approach, Sellers and Shmueli (2010) proposed a COM-Poisson regression model using the link function,

$$\eta(E(Y)) = \log \lambda = \mathbf{X}'\boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{p} \beta_j X_j.$$

Accordingly, this function indirectly models the relationship between $E(\mathbf{Y})$ and $\mathbf{X}'\boldsymbol{\beta}$, and allows for estimating $\boldsymbol{\beta}$ and ν via associated normal equations. Given the complexity of the normal equations, we solve them iteratively using the Poisson estimates, $\boldsymbol{\beta}^{(0)}$ and $\nu^{(0)} = 1$, as starting values. These equations can thus be solved via an appropriate iterative reweighted least squares procedure (or by maximizing the likelihood function directly using an optimization program) to determine the maximum likelihood estimates, $\hat{\boldsymbol{\beta}}$ and $\hat{\nu}$. The associated standard errors of the estimated coefficients are derived using the Fisher Information matrix; see Sellers and Shmueli (2010) and the accompanying online appendix for details. *R* code for estimating the COM-Poisson regression model is available on CRAN (`COMPoissonReg`). Note that this model assumes a constant dispersion level across all observations.

2.2. COM-POISSON REGRESSION WITH GROUP-LEVEL DISPERSION

We now introduce an extension of the COM-Poisson regression which allows for different

levels of dispersion across data subgroups. In particular, we use the link functions,

$$\log(\lambda) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j \tag{3}$$

$$\log(\nu) = \gamma_0 + \sum_{k=1}^{K-1} \gamma_k G_k \tag{4}$$

for the COM-Poisson parameters, where G_k is a dummy variable corresponding to one of the K groups in the data.

Estimating the coefficients, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, is done by maximum likelihood estimation, where the log-likelihood for observation i is given by

$$\log L_i(\lambda_i, \nu_i | y_i) = y_i \log \lambda_i - \nu_i \log y_i! - \log Z(\lambda_i, \nu_i) \tag{5}$$

with

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \doteq \mathbf{X}'_i \boldsymbol{\beta}, \text{ and} \tag{6}$$

$$\log(\nu_i) = \gamma_0 + \gamma_1 G_{i1} + \cdots + \gamma_{K-1} G_{i,K-1} \doteq \mathbf{G}'_i \boldsymbol{\gamma}. \tag{7}$$

Since the COM-Poisson distribution belongs to the exponential family, we can determine appropriate normal equations for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Using the Poisson estimates (i.e. $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\gamma}^{(0)} = \mathbf{0}$) as starting values, coefficient estimation can again be achieved via an appropriate iterative reweighted least squares procedure, or by using existing nonlinear optimization tools (e.g., `nlm` or `optim` in *R*) to directly maximize the likelihood function. The associated standard errors of the estimated coefficients are derived in an analogous manner to that described in Sellers and Shmueli (2010).

2.3. TESTING FOR VARIABLE DISPERSION

Sellers and Shmueli (2010) established a hypothesis testing procedure to determine if significant data dispersion exists, thus demonstrating the need for a COM-Poisson regression model over a simple Poisson regression model; in other words, they test whether $\nu = 1$ or otherwise. We now ask the follow-up question: is the dispersion level fixed across observations, or is it dependent on one or more of the $K - 1$ groups? More formally, we consider

the hypotheses,

$$\begin{aligned} H_0 &: \gamma_k = 0 \text{ for } k = 1, \ldots, K-1 \quad \text{vs.} \\ H_1 &: \gamma_k \neq 0 \text{ for at least one } k \in \{1, \ldots, K-1\}. \end{aligned} \tag{8}$$

The likelihood ratio statistic Λ and the derived test statistic C are given by

$$\Lambda = \frac{L\left(\hat{\boldsymbol{\beta}}_{(0)}, \hat{\boldsymbol{\gamma}}_{(0)}\right)}{L\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right)} \tag{9}$$

and

$$C = -2 \log \Lambda = -2 \left[\log L\left(\hat{\boldsymbol{\beta}}_{(0)}, \hat{\boldsymbol{\gamma}}_{(0)}\right) - \log L\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right)\right], \tag{10}$$

where $\hat{\gamma}_{k(0)} = 0$ for $k = 1, \ldots, K-1$. The maximum likelihood estimates obtained under H_0 are $\hat{\boldsymbol{\beta}}_{(0)}$ and $\hat{\boldsymbol{\gamma}}_{(0)}$, where $\boldsymbol{\nu}_{(0)} = \exp(\mathbf{X}\boldsymbol{\gamma}_{(0)})$; i.e., they are the COM-Poisson estimates under the constant dispersion model. Meanwhile, $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ are the maximum likelihood estimates under the variable dispersion model obtained by Equations (6) and (7). Under the null hypothesis, C has an approximate χ^2 distribution with $K-1$ degree of freedom, thus we reject H_0 in favor of H_1 when $C > \chi_{K-1}(\alpha)$. For small samples where the χ^2 approximation is questionable, a bootstrap procedure can be used.

3. EXAMPLE: ELEPHANT MATINGS

3.1. ORIGINATING DATA DESCRIPTION

Young adult male elephants must compete with older males to mate with receptive females. Because male elephants continue to grow in size throughout their lives, older elephants are larger and tend to be more successful at mating. Poole (1989) collected data on the age and number of successful matings for 41 male elephants in order to model the effect of age on the number of matings. The raw data were obtained from Ramsey and Schafer (2002).

3.2. POISSON AND COM-POISSON RESULTS

We first fit a Poisson regression model to the data, regressing the number of matings (Y) on age (X); the estimated coefficients are given by $\hat{\beta}_0 = -1.579, \hat{\beta}_1 = 0.069$. We then

fit a COM-Poisson regression model to determine whether equi-dispersion is a reasonable assumption. The resulting coefficients are close to those from the Poisson regression, the Pearson goodness-of-fit statistic is near 1, and the 90% bootstrap confidence interval for ν includes the value 1 (using 1,000 resamples); see Table I. Thus, it is reasonable to assume a Poisson regression to model this relationship.

Table I: Estimated regression models for elephant matings data.

	β_0	β_1	Dispersion
Poisson	-1.579	0.069	Pearson GOF=1.162
COM-Poisson	-1.448	0.060	$\hat{\nu}$=0.83 [90% CI =(0.53, 1.49)]

3.3. SIMULATED DATA

To illustrate the effect of mixing data with under- and over-dispersion, we simulated data based on the elephant matings dataset. In particular, given the age of the elephants in the dataset and the estimated β coefficients from the Poisson model (Table I), we generated the number of matings according to a COM-Poisson distribution. In Scenario 1, we generated 41 over-dispersed observations with $\nu_1 = 0.9$, and 41 under-dispersed observations with $\nu_2 = 2$, creating a mixed sample of 82 observations. In Scenario 2, we polarized the dispersion levels further, generating 41 over-dispersed observations with $\nu_1 = 0.9$ and 41 under-dispersed observations with $\nu_2 = 6$. Each of the two datasets is shown in Figure 1.

3.3.1. POISSON AND COM-POISSON REGRESSION

We start by fitting a Poisson regression and an ordinary COM-Poisson regression to the data. In particular, we estimate the models,

$$\log(\lambda) = \beta_0 + \beta_1 AGE \tag{11}$$

and

$$\log(\lambda) = \beta_0 + \beta_1 AGE + \beta_2 G, \tag{12}$$

where AGE denotes the elephant's age, and G is a dummy variable denoting whether the observation comes from the over-dispersed ($G = 1$) or under-dispersed ($G = 0$) group.

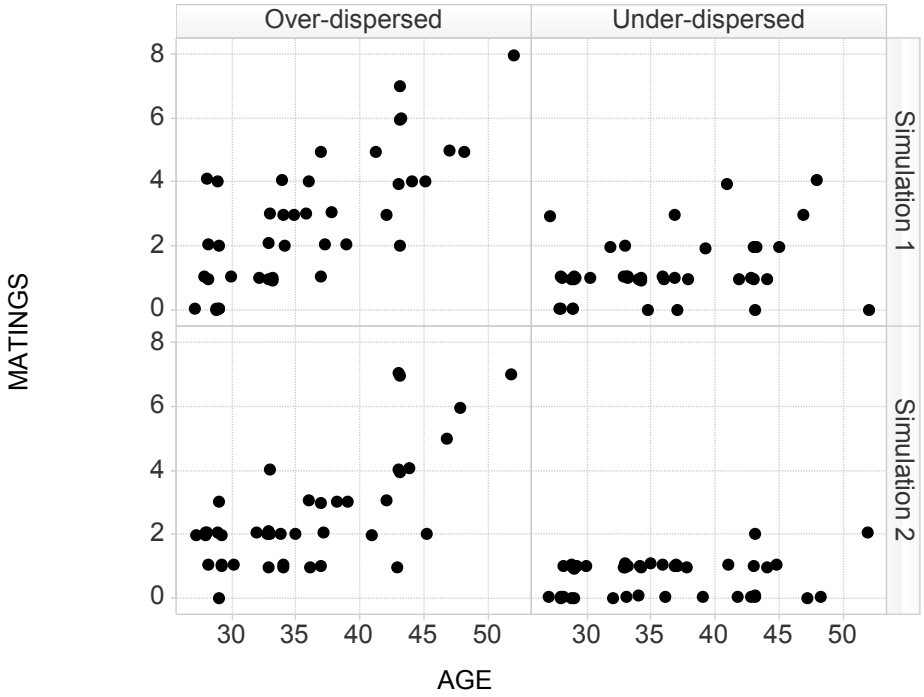

Figure 1: Scatterplot of Matings vs. Age for each of the two simulated datasets (top row contains Simulation 1; bottom row contains Simulation 2), by dispersion group (in columns). Slight jittering was used to avoid overlay of points.

Table II contains the parameter estimates for both simulations, considering the relationship between the number of matings solely in relation to age, and then with the added group effect. For Simulation 1, modeling the number of matings with age alone results in an apparent Poisson fit (as noted by the Pearson goodness of fit statistic for the Poisson regression and the 90% confidence interval for ν in the COM-Poisson regression; see also Figure 2). In contrast, for Simulation 2, the model indicates over-dispersion. In other words, the mixture of two dispersion levels can result in apparent equi-dispersion or over-dispersion even for different realizations of the same underlying structure.

Table II: Estimated model for simulated mixed matings data via various models, assuming fixed dispersion.

	Model	$\hat{\beta}_0$	$\hat{\beta}_1$ (Age)	$\hat{\beta}_2$ (Group)	Dispersion
Simulation 1	Poisson	-1.690	0.064	–	Pearson GOF=1.09
	COM-Poisson	-1.611	0.059	–	$\hat{\nu}$=0.89 [90% CI=(0.41, 1.02)]
	Poisson	-2.177	0.064	0.813	Pearson GOF=0.79
	COM-Poisson	-2.911	1.065	0.092	$\hat{\nu}$=1.462 [90% CI=(0.81,1.68)]
Simulation 2	Poisson	-1.646	0.057	–	Pearson GOF=1.13
	COM-Poisson	-1.508	0.047	–	$\hat{\nu}$=0.72 [90% CI=(0.19, 0.63)]
	Poisson	-2.577	0.057	1.405	Pearson GOF=0.55
	COM-Poisson	-3.141	2.133	0.082	$\hat{\nu}$=1.90 [90% CI=(0.81,1.83)]

For both simulations, adding the group effect as an additional covariate in the model results in *apparent underdispersion* according to the Pearson statistic (< 1), and to *apparent equi-dispersion* according to the 90% confidence intervals for ν and as can be seen in the bottom panels of Figure 2. In both cases the real dispersion structure is disguised.

In the next section, we will see that taking into account the group assignment by relating it to the dispersion parameter helps detect the difference in group-level dispersion levels. This is a significant issue because an analyst who considers only the results of linking group membership to λ is likely to conclude that Poisson or negative binomial regression (respectively) are adequate models when, in fact, they cannot capture the underlying dispersion structure.

3.3.2. COM-POISSON REGRESSION WITH GROUP-LEVEL DISPERSION

We now examine each of the two simulated datasets by fitting a COM-Poisson regression model with group-level dispersion, using the link equations,

$$\log(\lambda) = \beta_0 + \beta_1 AGE \tag{13}$$

and

$$\log(\nu) = \gamma_0 + \gamma_1 G. \tag{14}$$

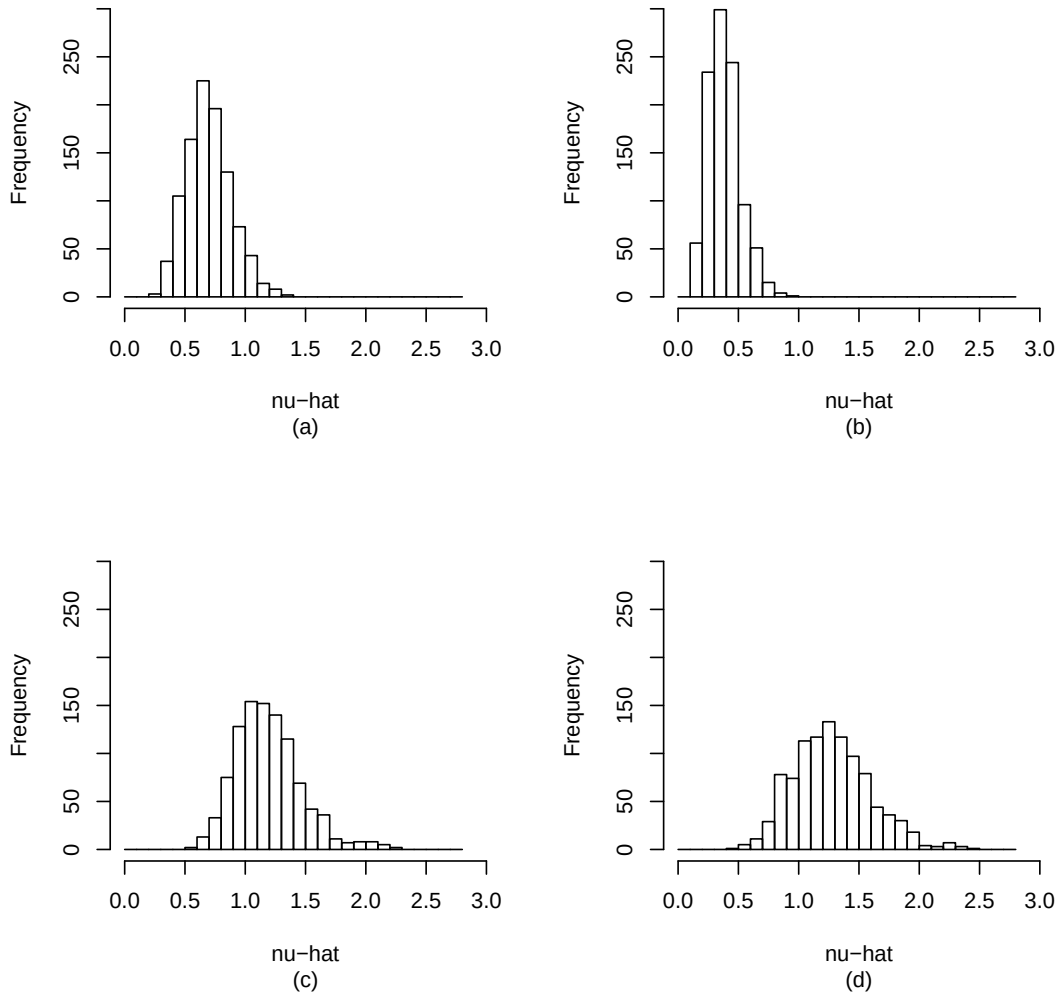

Figure 2: Histogram of $\hat{\nu}$ for each of the two simulated datasets (left = Simulation 1, right = Simulation 2), based on the 1,000 bootstrap samples. Top panels: AGE as a single covariate. Bottom panels: AGE and G as two covariates.

The estimated models for each of the simulated datasets are given in Table III.

Table III: Estimated COM-Poisson models with group-level dispersion (Equations (13)-(14)) for simulated mixed matings data.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}_0$	$\hat{\gamma}_1$
Simulation 1				
True parameters	-1.579	0.069	log(2) = 0.693	log(0.9/2)= -0.799
Group-level dispersion	-1.843	0.080	0.828	-0.725 [90% CI=(-1.044, -0.602)]
Simulation 2				
True parameters	-1.579	0.069	log(6) = 1.792	log(0.9/6)= -1.897
Group-level dispersion	-1.730	0.075	1.692	-1.561 [90% CI=(-4.632, -1.557)]

The test to show the existence of a statistically significant mixture of dispersion levels (i.e. $H_0 : \gamma_1 = 0$ versus $H_1 : \gamma_1 \neq 0$) yields respective test statistics of 29.928 and 59.250 with associated p-values ≈ 0, indicating (in fact) that a mixture of dispersion levels exists. Due to the small sample size, we confirmed this result by using a 90% bootstrap confidence interval for γ_1 (based on 1,000 resamples). In both cases, the interval did not contain 0. In summary, the group-level COM-Poisson model was able to detect the mixture of dispersion levels rather than treat the entire sample as coming from a single population mistakenly presumed to be either equi- or over-dispersed. Moreover, the fitted model produces estimates that are quite close to the true underlying parameters used to generate the data.

Finally, to assess the ability of the group-level dispersion COM-Poisson model to correctly identify that groups differ only in dispersion level but not in λ, we fit the model,

$$\log(\lambda) = \beta_0 + \beta_1 AGE + \beta_2 G \tag{15}$$

$$\log(\nu) = \gamma_0 + \gamma_1 G. \tag{16}$$

The results in Table IV show that, for both simulations, the model identifies that the groups differ only in dispersion level ($\beta_2 = 0$ and $\gamma_1 \neq 0$).

Table IV: 90% bootstrap confidence intervals (based on 1,000 resamples) for β_2 and γ_1 fitting a model with Equations (15)-(16), for each of the simulations.

	β_2	γ_1
Simulation 1	(-0.761, 0.756)	(-1.330, -0.308)
Simulation 2	(-0.722, 0.886)	(-5.468, -1.341)

3.4. THREE GROUP EXAMPLE

To illustrate the generality of the apparent dispersion effect when mixing more than two groups and to further illustrate apparent dispersion, we combine the simulated data (which mix over- and under-dispered data) with the original equi-dispersed data. Figure 3 shows scatterplots for each of the three groups (in columns), separately for each of the two simulated datasets (top and bottom panels, correspondingly). In the following models, we use two groups dummy variables (*Group1*=1 if over-dispersed, *Group2*=1 if equi-dispersed) with the under-dispersed group serving as the reference category, and fit the Poisson, COM-Poisson with fixed dispersion, and COM-Poisson with group-level dispersion models.

Fitting a COM-Poisson regression model to each of Simulations 1 and 2 separately, and examining the estimated ν parameter, we see that Simulation 1 displays apparent equi-dispersion, while Simulation 2 displays apparent over- or under-dispersion, depending on the factors included in the model; see Table V. Apparent equi-dispersion exists whether or not the group factors are included in the model for Simulation 1, while the inclusion of these group factors impacts the apparent form of dispersion detected in Simulation 2. This apparent dispersion is also evident from the Pearson GOF statistic from a fitted Poisson regression model. The results seen here are consistent with the two-group mixing that we saw earlier, as demonstrated by the Pearson statistic from the Poisson models and the 90% confidence intervals for ν from the COM-Poisson model with fixed ν. In short, as in the two-group mixture, dispersion is disguised in the three-group mixture.

Next, we fit a COM-Poisson regression with group-level dispersion. The three groups are incroporated into the model using two dummy variables as group covariates in each of Equations (6)-(7). The results are shown in Table VI. The first model, where dispersion is

Table V: Estimated model for three-group mixed matings data via various models, assuming fixed dispersion.

Model	$\hat{\beta}_0$ (Intercept)	$\hat{\beta}_1$ (Age)	$\hat{\beta}_2$ (Group1)	$\hat{\beta}_3$ (Group2)	Dispersion
Simulation 1					
Poisson	-1.655	0.066	–	–	Pearson GOF=1.268
COM-Poisson	-1.521	0.057	–	–	$\hat{\nu}$=0.81 [90% CI=(0.605, 1.130)]
Poisson	-2.245	0.813	0.769	0.066	Pearson GOF=1.0334
COM-Poisson	-2.416	0.895	0.845	0.073	$\hat{\nu}$= 1.145 [90% CI=(0.912, 1.565)]
Simulation 2					
Poisson	-1.647	0.062	–	–	Pearson GOF=1.4179
COM-Poisson	-1.424	0.046	–	–	$\hat{\nu}$=0.63 [90% CI=(0.428, 0.949)]
Poisson	-2.779	1.405	1.442	0.062	Pearson GOF=0.8765
COM-Poisson	-3.266	1.696	1.743	0.079	$\hat{\nu}$=1.354 [90% CI=(1.071, 1.885)]

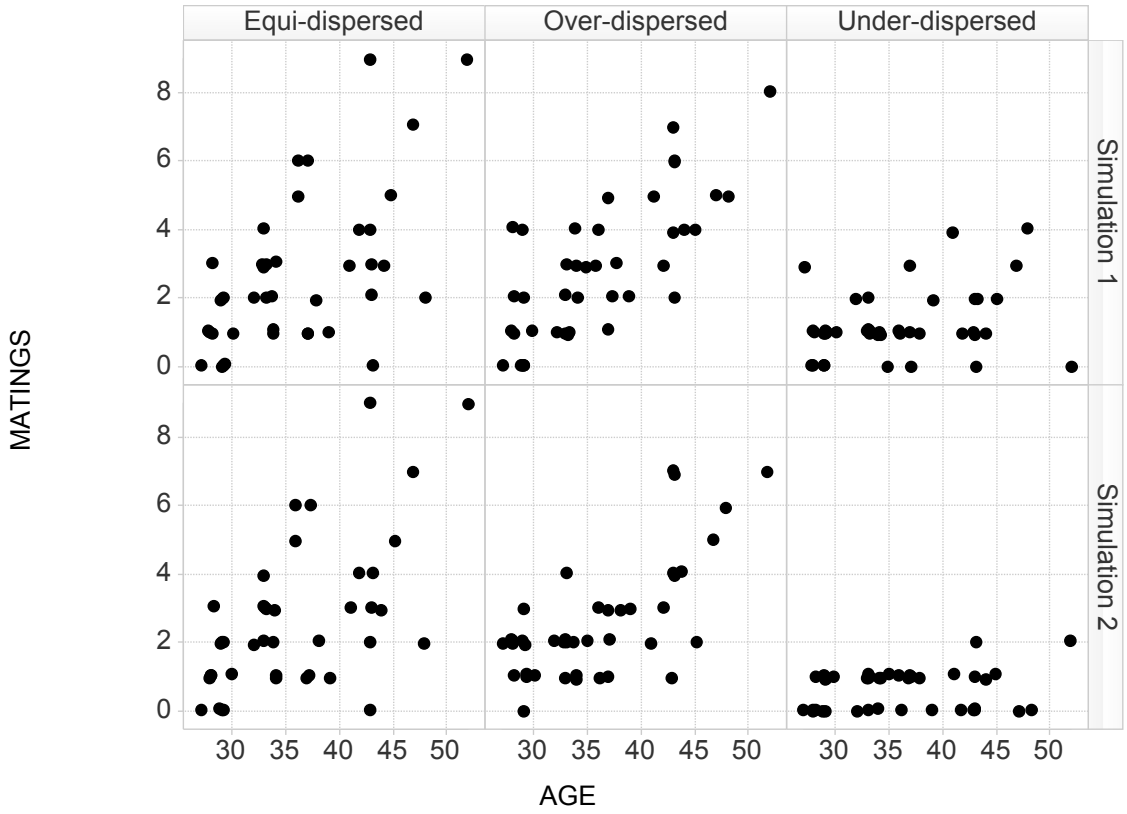

Figure 3: Scatterplot of Matings vs. Age for each of the two simulated datasets (top= Simulation 1, bottom = Simulation 2), by dispersion group (columns 1-3). Slight jittering was used to avoid overlay of points.

incorporated only into the equation for ν, accurately captures the three levels of dispersion. In particular, we see that $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are statistically significant. Moreover, the estimates are very close to the true parameters from which they were simulated (true also for the β estimates). In fact, Simulation 1 further shows that adding the group variables into the equation for λ still results in $\hat{\gamma}_1$ and $\hat{\gamma}_2$ being significant while $\hat{\beta}_2$ and $\hat{\beta}_3$ are not so. Thus, we see that what first appears as a Poisson dataset is actually a mixture of subgroups with varying levels of dispersion.

4. CONCLUDING REMARKS

Identifying and fitting a mixed-dispersion dataset requires a model that can capture both

Table VI: Estimated COM-Poisson models with group-level dispersion (Equations (6)-(7)) for 3-group mixed matings data. Confidence intervals based on 100 simulations via parametric bootstrap.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\gamma}_2$
	$\log(\lambda)$ Int.	Age	Group 1	Group 2	$\log(\nu)$ Int.	Group 1	Group 2
Simulation 1							
True Parameters	-1.579	0.069	-	-	0.693	-0.799	-0.693
Group-level disp.	-1.667	0.071	-	-	0.741	-0.747	-0.737
90% CI	(-2.476, -1.187)	(0.059, 0.097)	-	-	(0.50, 1.124)	(-0.996, -0.517)	(-0.955, -0.552)
Group-level disp.	-1.836	0.073	0.304	0.052	0.673	-0.561	-0.710
90% CI	(-2.856, -0.863)	(0.053, 0.097)	(-0.315, 1.127)	(-0.814, 0.763)	(0.262, 1.105)	(-1.099, -0.016)	(-1.274, -0.193)
Simulation 2							
True Parameters	-1.579	0.069	-	-	1.792	-1.897	-1.792
Group-level disp.	-1.594	0.068	-	-	1.653	-1.605	-1.688
90% CI	(-2.382,-0.941)	(0.045,0.095)	-	-	(1.385,4.062)	(-4.034,-1.292)	(-4.115,-1.384)
Group-level disp.	-2.507	0.078	1.234	0.577	1.457	-0.998	-1.435
90% CI	(-3.591, -1.325)	(0.056, 0.110)	(0.355, 2.215)	(-0.108, 1.499)	(1.137, 4.986)	(-4.723, -0.462)	(-4.961, -0.897)

over- and under-dispersion. Although the negative binomial regression model is very popular for modeling over-dispersed data, if the over-dispersion is a guise for an underlying mixture of dispersion levels, it is conceptually and practically more appealing to use the COM-Poisson regression described here.

Mixtures of strictly over-dispersed (or strictly under-dispersed) populations generally produce an overall dataset that is likewise over-dispersed (or under-dispersed). Although one might expect the mixed dataset to reflect a dispersion level that falls between the dispersion levels of the different groups, we have found that mixtures of over-disperesed groups can result in an overall over-dispersion level that is even more extreme than each of the separate groups. Hence, apparent over-dispersion can actually be a disguise for a mixture of groups each having lower over-dispersion levels. We further found that mixing different types of dispersion (i.e. over-, equi-, and under-dispersed data) can likewise appear over-, equi-, or even under-dispersed! Hence, the apparent dispersion based on a Poisson regression GOF statistic or even on the estimated ν in a fixed-ν COM-Poisson regression model can be misleading.

We introduce here an extension to the COM-Poisson regression model that can detect and capture data that come from mixtures of dispersion levels. Our focus was on group-level

dispersion, assuming that we have multiple observations coming from each dispersion level. Further we consider the example of a simulated dataset so that we can accurately assess the impact made in properly accounting for group dispersion in the model. Meanwhile, a related model for consideration is one where the dispersion level is observation-specific, that is, a model with link functions of the form,

$$\log(\lambda_i) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j} \tag{17}$$

$$\log(\nu_i) = \gamma_0 + \sum_{k=1}^{q} \gamma_k V_{k,i}, \tag{18}$$

where the covariates used in each of the equations can either differ or overlap. Such models are especially natural in marketing, where customer-level heterogeneity is of interest. For examplem, Boatwright et al. (2003) describe the importance of capturing customer heterogeneity in eCommerce data via the parameters of the statistical model:

> "The analyst of grocery data faces [the] challenge to account for the heterogeneity of customers. Historically, many researchers have attempted this by incorporating customer level information such as demographics. Even though some online retailers request such information from their customers (household size, income), the data is typically noisy as households are increasingly reluctant to divulge personal information... An alternative is to estimate models with customer-specific parameters, where household purchase histories provide the information on customer heterogeneity."

For further details on an observation-level COM-Poisson model, see Sellers and Shmueli (2009). Additional models can be derived along these lines. For example, taking a Bayesian approach, one could supplement the observation-level model with a distribution on ν to create group-level dispersion. The choice between a group-level, observation-level, or other model would depend on knowledge about the underlying mechanism producing the data. In this paper, we focused on the common case of data arising from mixtures of observations from populations with different dispersion levels.

BIBLIOGRAPHY

Boatwright, P., Borle, S., and Kadane, J. B. (2003). A Model of the Joint Distribution of Purchase Quantity and Timing. *Journal of the American Statistical Association*, **98**(463), 564–571.

Borle, S., Dholakia, U., Singh, S., and Westbrook, R. (2007). The impact of survey participation on subsequent behavior: an empirical investigation. *Marketing Science*, **26**, 711-726.

Castillo, J., and Prez-Casany, M. (1998). Weighted Poisson Distributions for Overdispersion and Underdispersion Situations. *Ann. Inst. Statist. Math.*, **50**, 567–585.

Castillo, J., and Prez-Casany, M. (2005). Overdispersed and Underdispersed Poisson Generalizations. *Journal of Statistical Planning and Inference*, **134**(2), 486–500.

Famoye, F. (1993). Restricted Generalized Poisson Regression Model. *Communications in Statistics - Theory and Methods*, **22**, 1335–1354.

Hilbe, J. M. (2007). *Negative Binomial Regression*. Cambridge University Press, 5th edition.

Kokonendji, C. C., and Mizere, D. and Balakrishnan, N. (2008). Connections of the Poisson Weight Function to Overdispersion and Underdispersion. *Journal of Statistical Planning and Inference*, **138**(5), 1287 – 1296.

McLaren C. E., Wagstaff, M., Brittenham, G. M., and Jacobs, A. (1991). Detection of Two-Component Mixtures of Lognormal Distributions in Grouped, Doubly Truncated Data: Analysis of Red Blood Cell Volume Distributions. *Biometrics*, **47**(2), 607–622.

Park, B. J., and Lord, D. (2009). Application of Finite Mixture Models for Vehicle Crash Data Analysis. *Animal Behaviour*, **41**(4), 683–691.

Poole, J. (1989). Mate Guarding, Reproductive Success and Female Choice in African Elephants. *Animal Behaviour*, **37**, 842–849.

Ramsey, F. and Schafer, D. (2002). *The Statistical Sleuth: A Course in Methods of Data*

Analysis. Duxbury Press, 2nd edition.

Sellers, K. F. and Shmueli, G. (2009). A Regression Model for Count Data with Observation-Level Dispersion. In Booth, J. G., editor, *Proceedings of the 24th International Workshop on Statistical Modelling*, Ithaca, NY, 337–344.

Sellers, K. F. and Shmueli, G. (2010). A Flexible Regression Model for Count Data. *Annals of Applied Statistics*, **4**, 943–961.

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. *Applied Statistics*, **54**, 127–142.