Data Dispersion: Now You See it... Now You Don't

Kimberly F. Sellers

&

Galit Shmueli

http://eprints.exchange.isb.edu/296/

Working Paper

Indian School of Business

2010

Data Dispersion: Now You See It... Now You Don't

Kimberly F. Sellers

Department of Mathematics

Georgetown University, Washington, DC 20057

Galit Shmueli

Department of Decision, Operations & Information Technologies

Smith School of Business, University of Maryland, College Park, MD 20742

Abstract

The most popular method for modeling count data is Poisson regression. When data display over-dispersion, thereby deeming Poisson regression inadequate, typically negative-binomial regression is instead used. We show that count data that appear to be equi-dispersed 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 that allows for group-level dispersion. We illustrate mixed dispersion effects and the proposed methodology via semi-authentic data.

Keywords: Conway-Maxwell-Poisson (COM-Poisson) regression, dispersion, mixture model, negative binomial regression, under-dispersion

1 Introduction

Poisson regression is a popular tool for modeling count data, however, it is limiting in its assumption that the population variance equals the expected value. Thus, it is not an appropriate regression technique for most real-world data as they are usually over- or under-dispersed. Overdispersion is a more frequent issue for a variety of reasons, e.g. excessive zero counts, censoring, etc; as a result, it is a focus of interest among statisticians. In this case, the negative binomial regression is a popular choice used to model the relationship between the predictor and explanatory variables. Its popularity is illustrated through associated regression tools being widely available in many statistical software packages, and is generally accepted as a reasonable method for modeling overdispersed count data.

In the event that the data is actually underdispersed, however, the negative binomial regression is not equipped to model such data. Other approaches exist that allow for over- or under-dispersion. The restricted generalized Poisson regression of Famoye (1993), for example, has the form

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

$$y_i = 0, 1, 2, \ldots,$$

where $\alpha = 0$ represents the special case of a Poisson regression; $\alpha > 0$ handles overdispersion, and $-2/\mu_i < \alpha < 0$ addresses data underdispersion. This parameter, however, is notably restricted in the amount of underdispersion that it can model; see Famoye (1993) for discussion. Meanwhile, the Conway-Maxwell-Poisson (COM-Poisson) regression of Sellers and Shmueli (2010) is a regression tool that generalizes Poisson regression, and can successfully model count data containing any sort of dispersion.

All regression models that accommodate for data dispersion stem from an underlying distribution whose mean is a function of both parameters. This is in stark contrast to a normal model, for example, where the parameters (μ and σ) are independent. This matter is significant particularly when one considers mixtures of these distributions and the resulting impact of such a data structure on the regression in question. In the normal distribution, if populations

with the same mean but different variances are mixed, the result will be a normal distribution with the same mean and a variance that is a "compromise" between the individual variances. In contrast, the result of mixing over- and under-dispersed count data can lead to data that appear to have any sort of dispersion.

Given a dataset, how can one determine whether its associated dispersion (or lack thereof) is "real" or masked when it really results from a mixture of dispersions? In this paper, we propose a COM-Poisson regression model that allows for variable dispersion to address this problem. The model can then be used to test for different levels of dispersion across different groups, and to estimate the associated respective dispersion levels within each group.

Hilbe (2007) addresses the matter of apparent versus real overdispersion, arguing that apparent overdispersion occurs as a result of one or more of the following: the model omitting important explanatory predictors, outliers in the data, model consideration that does not include enough interaction terms, misrepresentation of a predictor variable in the appropriate scale, or misspecification of the link function. This implies that correcting for any of the above matters could produce an equidispersed dataset that can be fit via Poisson regression. In this paper, we introduce a new important factor impacting apparent data dispersion, namely mixtures of dispersion levels. Data resulting from such a mixture can display apparent dispersion of any form, when the underlying structure mixes different levels of dispersion (even a mixture of over- and under-dispersion, as illustrated here).

The paper is organized as follows. Section 2 briefly describes the COM-Poisson distribution and regression model, which assume a constant 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 a semi-authentic dataset on elephants matings, where a sample of over-dispersed data are 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 The COM-Poisson distribution

The COM-Poisson probability mass function takes the form

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

for a random variable Y_i, where $Z(\lambda_i, \nu) = \sum_{s=0}^{\infty} \frac{\lambda_i^s}{(s!)^\nu}$. In this setting, $\lambda_i = E(Y_i^\nu)$, while ν 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.

2.2 COM-Poisson regression with constant dispersion

Taking a GLM approach, Sellers and Shmueli (2010) proposed a COM-Poisson regression model using the link function $\eta(E(Y)) = \log \lambda$. This function indirectly models the relationship between $E(\mathbf{Y})$ and $\mathbf{X}\boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$, and allows for estimating $\boldsymbol{\beta}$ and ν via associated normal equations. Using the Poisson estimates, $\boldsymbol{\beta}^{(0)}$ and $\nu^{(0)} = 1$ as starting values, these equations can be solved iteratively via an appropriate iterative reweighted least squares procedure (or by maximizing the maximum likelihood directly using an optimization program) to determine the maximum likelihood estimates for $\boldsymbol{\beta}$ and ν. 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 at `www9.georgetown.edu/faculty/kfs7/research`. Note that this model assumes a constant dispersion level across all observations.

2.3 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 different groups in the data. In particular, we use respective link functions

for the COM-Poisson parameters, namely:

$$\log(\lambda) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j \tag{2}$$

$$\log(\nu) = \gamma_0 + \sum_{k=1}^{K-1} \gamma_k G_k, \tag{3}$$

where G_k is a dummy variable corresponding to one of the K groups in the data.

Estimating the β and γ coefficients is done by maximizing the log-likelihood. The log-likelihood for observation i is given by

$$\log L_i(\lambda_i, \nu_i | y_i) = y_i \log \lambda_i - \nu_i \log y_i! - \log Z(\lambda_i, \nu_i), \tag{4}$$

where

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \doteq \mathbf{X}_i \boldsymbol{\beta}, \text{ and} \tag{5}$$

$$\log(\nu_i) = \gamma_0 + \gamma_1 G_{i1} + \cdots + \gamma_{K-1} G_{i,K-1} \doteq \mathbf{G}_i \boldsymbol{\gamma}. \tag{6}$$

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, $\boldsymbol{\beta}^{(0)}$ 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).

2.4 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 set of hypotheses:

$$\begin{aligned} H_0 &: \gamma_k = 0 \text{ for } k = 1, \ldots, K-1 \quad \text{vs.} \\ H_1 &: \gamma_k \neq 0 \text{ for at least one } k \in \{1, \ldots, K-1\}. \end{aligned} \tag{7}$$

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

$$\Lambda = \frac{L\left(\hat{\boldsymbol{\beta}}_{(0)}, \hat{\boldsymbol{\gamma}}_{0(0)}\right)}{L\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right)} \tag{8}$$

$$C = -2\log\Lambda = -2\left[\log L\left(\hat{\boldsymbol{\beta}}_{(0)}, \hat{\boldsymbol{\gamma}}_{0(0)}\right) - \log L\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right)\right], \tag{9}$$

where $\hat{\gamma}_{i(0)} = 0$ for $j = 1, \ldots, K-1$. $\hat{\boldsymbol{\beta}}_{(0)}, \hat{\boldsymbol{\gamma}}_{(0)}$ where $\nu_{(0)} = \exp(\mathbf{X}\boldsymbol{\gamma}_{(0)})$ are the maximum likelihood estimates obtained under H_0, that is, they are the COM-Poisson estimates under the constant dispersion model; and $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ are the maximum likelihood estimates under the variable dispersion model, obtained by Equations (5) and (6). Under the null hypothesis, C has an approximate χ^2 distribution with $K-1$ degree of freedom. Therefore, 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 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-

Table 1: 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))

Poisson regression model to determine whether equi-dispersion is a reasonable assumption. The resulting coefficients are close to those from the Poisson regression (see Table 1), 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). Hence, it is reasonable to assume a Poisson regression to model this relationship.

3.3 Simulated data

To illustrate the effect of mixing data with under- and over-dispersion, we simulated data based on the elephants matings dataset. In particular, given the age of the elephants in the dataset and the estimated β coefficients from the Poisson model (Table 1, second row), 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. Note that the under-dispersed group is concentrated at the bottom of each plot (triangles).

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 two models:

$$\log(\lambda) = \beta_0 + \beta_1 AGE \tag{10}$$

and

$$\log(\lambda) = \beta_0 + \beta_1 AGE + \beta_2 G, \tag{11}$$

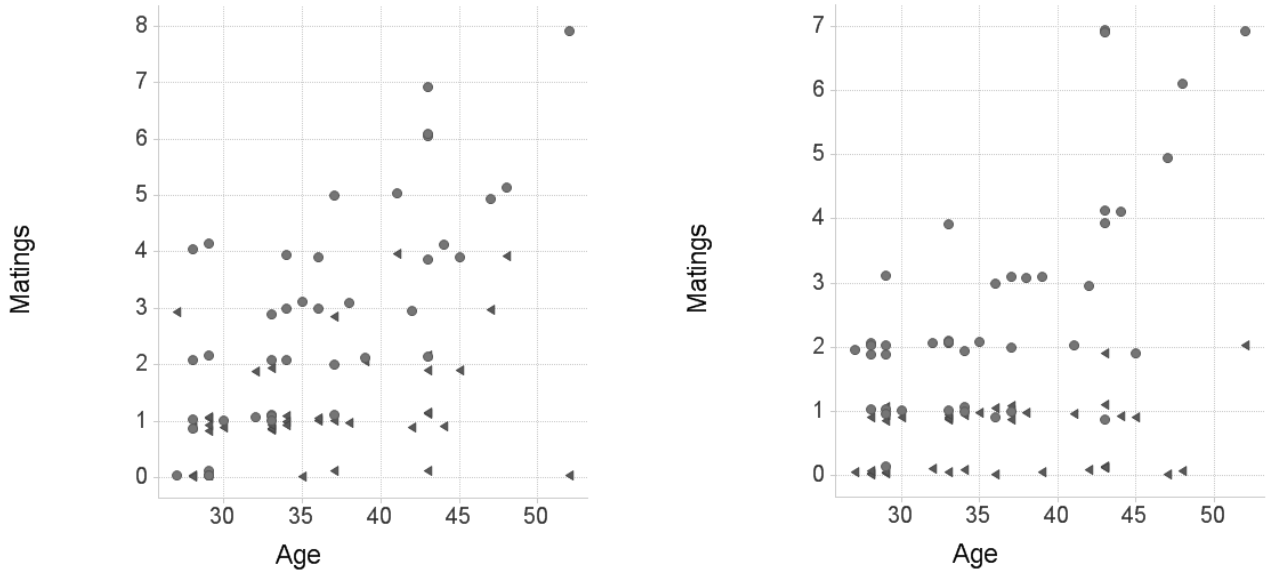

Figure 1: Scatterplot of Matings vs. Age for each of the two simulated datasets (left = Simulation 1, right = Simulation 2). Triangles denote the under-dispersed group, and circles denote the over-dispersed group. Slight jittering was used to avoid overlay of points.

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. Table 2 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, both models indicate 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.

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 mem-

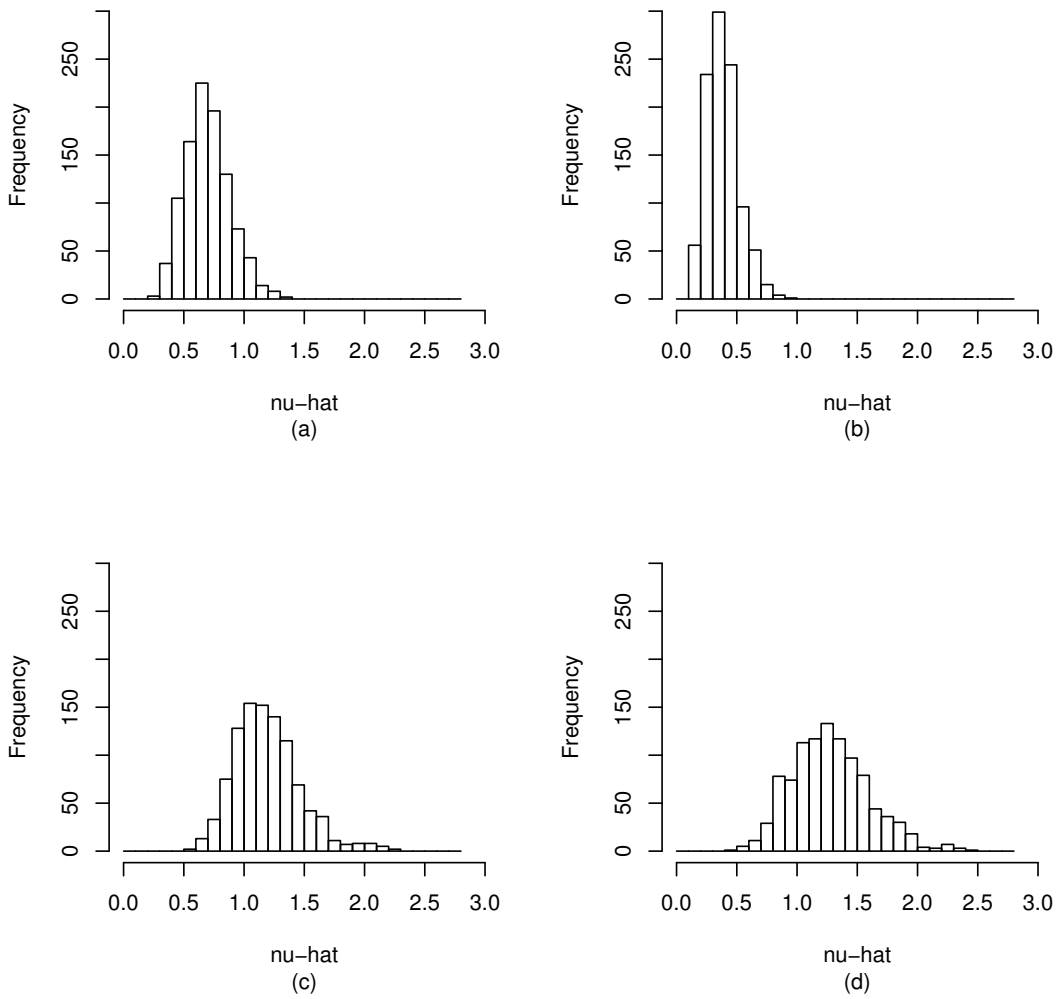

Figure 2: Histogram of $\hat{\nu}$ for each of the two simulated datasets (left = Simulation 1, right = Simulation 2), based on the 1,000 bootstrap samples. Top panels: AGE as a single covariate. Bottom panels: AGE and G as two covariates.

Table 2: Estimated model for simulated mixed matings data via various models

	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))

Table 3: Estimated COM-Poisson models with group-level dispersion (Equations (12)-(13)) for simulated mixed matings data.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}_0$	$\hat{\gamma}_1$
Simulation 1				
True estimates	-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 estimates	-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))

bership 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 following link equations:

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

$$\log(\nu) = \gamma_0 + \gamma_1 G. \tag{13}$$

The estimated models for each of the simulated datasets are given in Table 3.

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

Table 4: 90% bootstrap confidence intervals (based on 1,000 resamples) for β_2 and γ_1 fitting a model with equations (14)-(15), 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)

identify that groups differ only in dispersion level but not in λ, we fit the model

$$\log(\lambda) = \beta_0 + \beta_1 AGE + \beta_2 G \quad (14)$$

$$\log(\nu) = \gamma_0 + \gamma_1 G. \quad (15)$$

The results in Table 4 show that for both simulations the model identifies that the groups differ only in dispersion level ($\beta_2 = 0$ and $\gamma_1 \neq 0$).

4 Concluding Remarks

Identifying and fitting a mixed-dispersion dataset requires a model that can capture both 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 by Sellers and Shmueli (2010).

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 be disguising a mixture of groups which each has a lower over-dispersion level.

We introduced 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. A related model is one where the dispersion level is observation-specific, i.e. a model with link

functions of the form,

$$\log(\lambda_i) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j} \tag{16}$$

$$\log(\nu_i) = \gamma_0 + \sum_{k=1}^{q} \gamma_k V_{k,i}, \tag{17}$$

where the covariates used in each of the equations can either differ or overlap. 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.

References

Famoye, F. (1993). Restricted generalized Poisson regression model. *Communications in Statistics - Theory and Methods*, 22(5):1335–1354.

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

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, 2 edition.

Sellers, K. 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*, pages 337–344, Ithaca, NY.

Sellers, K. and Shmueli, G. (2010). A flexible regression model for count data. *Annals of Applied Statistics*, forthcoming.

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.