Hacettepe Journal of Mathematics and Statistics
Volume 47 (5) (2018), 1335 – 1347

Overdispersed count models for mRNA transcription

Burcin Simsek*† and Satish Iyengar‡

Abstract

Direct detection of gene activity is often not possible because new proteins from an individual activation event are masked by proteins remaining from previous events. Thus, researchers determine gene activation or inactivation by observing messenger RNA (mRNA) production instead. Typically, mRNA transcription occurs in short rapid bursts when the gene is in its on-state, and no transcriptions during its off-state. This burstiness of mRNA production is not well modeled by a Poisson process. We propose the Conway-Maxwell-Poisson (COM-Poisson) distribution as a potential alternative to the more common negative binomial (NB) distribution. We use the generalized linear model version of these models to incorporate covariate information. We also consider zero inflation to model excess zero counts. We use data from *E. coli* bacteria and mammalian cells to illustrate our proposed methods. We find that when there is a biophysically derived distribution, this distribution performs well. We also show that in the absence of such biophysical knowledge, the COM-Poisson is competitive with the NB. Both the COM-Poisson and NB arise in queueing theory, suggesting that further application of that framework to study mRNA dynamics would be useful.

Keywords: Conway-Maxwell-Poisson, Link function, Model comparison, Negative binomial, Generalized linear model.

Mathematics Subject Classification (2010): 62J12, 62P10, 92C99

Received : 06/09/2016 *Accepted* : 19/01/2017 *Doi* : 10.15672HJMS.2017.422

*Department of Statistics, University of Pittsburgh, Pittsburgh, USA, Email: bus5@pitt.edu
†Corresponding Author.
‡Department of Statistics, University of Pittsburgh, Pittsburgh, USA, Email: ssi@pitt.edu

1. Introduction

Variations in the nature of the biochemistry of gene expression lead to variations in protein levels. These variations in turn lead to variations in cell function, even among cells of similar genotype. The consequences of such variations that lead to phenotypic heterogeneity are discussed in [6]. These variations appear to have a stochastic component, so it is of interest to understand their nature. Direct measures of gene activity using measurements of proteins are difficult because of masking from previous events [16]. However, an indirect measure is available through the protein's precursor called messenger RNA (mRNA), a molecule that is a transcription (copy) of a segment of DNA known as the promoter. Variations in protein synthesis have been attributed to variations in mRNA transcription. Measurement of mRNA is feasible and less susceptible to masking. It also serves as an indirect indicator of gene activity; thus, there is considerable interest in the study of mRNA counts [5]. Quantification of mRNAs is done through the use of fluorescence *in situ* hybridization (FISH), with mRNA molecules in each cell being counted by identifying individual fluorescent spots of 3-dimensional images [16, 24, 25].

Two prominent models proposed for mRNA transcription are Markovian: the Poisson and the telegraph processes, both of which are commonly used in queueing theory [5, 8, 27]. The telegraph model is a two-state Markov process with the two states representing the active and inactive phases of the promoter. A notable feature of mRNA transcription is its burstiness, which appears to lend more support to the telegraph model [14, 24]. It also suggests that counts would be overdispersed – that is, the variance is larger than the mean – relative to the Poisson model: see the Appendix. However, there are attempts to explain the burstiness under both models. In addition, it appears that the protein bursts themselves do not uniquely identify mRNA burst distributions.

There have been several studies of mRNA transcription and other biological network processes from a queueing theory perspective [1, 5]. The analogy with queueing theory is as follows: the times of mRNA transcriptions correspond to arrival times of customers whose waiting times in the queue correspond to the time to degradation of the mRNA molecule. This analogy allows the considerably large body of research from queueing theory into this context. For example, see [5] for an application of Little's law, which relates the burst and steady state means of mRNA production. However, such investigations are still incomplete: although in certain circumstances master equations are known which lead to models that fit well (see Section 3.2 below), in many other cases models based on such detailed knowledge are not available. Thus, further studies which could lead to a better understanding of commonalities across a wide range of mRNA transcription data are necessary. Earlier studies [20, 24] have modeled the count distribution using the negative binomial (NB) distribution as an alternative to the Poisson distribution. However, the Conway-Maxwell Poisson distribution (COM-Poisson, [4, 17, 21]) has not yet been applied to mRNA counts. We expect that the COM-Poisson will be a good candidate to model the mRNA transcription's burstiness because it allows overdispersion; also, it offers an another queueing model to consider.

The details of the experiments involving mRNA counts are important. Thus, the count models should be extended so that they allow the estimation of the effects of experimental conditions through appropriate covariates. In short, regression models for the counts are necessary. GLM methods for mRNA counts were suggested by Zhang, et al. [28], but they do not appear to have been studied systematically. Thus, our aim in this paper is to apply COM-Poisson regression models [3, 9, 11, 19] to overdispersed mRNA count data and to compare their performance with other well known models for count data. We also use the zero-inflated COM-Poisson (ZICOM-Poisson) regression model [2, 18] to handle excessive number of zeroes when they occur.

In Section 2, we present our notation for the NB, Poisson, a biophysically based model (discretized gamma), and the COM-Poisson distribution. We also use their GLM versions and their zero-inflated versions. Because the NB and Poisson models are well known, we only present our notation for them. In Section 3, we apply the proposed models to *E. coli* and mammalian cell mRNA count data under different experimental conditions. We then compare the model fits using the Bayes information criterion (BIC) and provide a few parameter estimates and model diagnostics as illustrative examples. In Section 4, we conclude with a discussion of the implications of our work and possible avenues for the future. Finally, in the Appendix, we use a doubly stochastic point process to show how bursty behavior can lead to overdispersion.

2. Regression models for count data

The regression models that we consider are the COM-Poisson, Poisson, NB, and a biophysically based model. We call the probability mass function (pmf) of this biophysical model a discretized gamma because it resembles a gamma density. Developments for the COM-Poisson are relatively recent, so we give a brief account of it and key references. Although the Poisson is a special case of COM-Poisson, we present results for it separately because it is a widely used regression model for count data. The NB is also well known, so we just present our notation for it.

2.1. COM-Poisson regression model. The COM-Poisson was introduced by Conway and Maxwell [4] for modeling queues and service rates. It has recently been shown to be useful in many other applications, especially by Shmueli, Sellers, and their colleagues [17, 21]. This distribution is a two-parameter discerete distribution on $\{0, 1, 2, \ldots\}$. Its parameters λ and ν model the intensity and the dispersion, respectively. For $k = 0, 1, 2, \ldots$, its pmf is

$$P(X = k|\lambda, \nu) = Z(\lambda, \nu)^{-1} \frac{\lambda^k}{(k!)^\nu}, \quad \text{where} \quad Z(\lambda, \nu) = \sum_{k=0}^{\infty} \frac{\lambda^k}{(k!)^\nu},$$

and its parameter space is

$$\Theta = \{(\lambda, \nu) : \lambda > 0, 0 < \nu \leq \infty\} \cup \{(\lambda, \nu) : 0 < \lambda < 1, \nu = 0\}.$$

The normalizing constant $Z(\lambda, \nu)$ does not have an easy closed form expression. Thus, there have been several recent studies of approximations for it [7, 12, 15, 22, 23]. The distribution is overdispersed (underdispersed) if $0 \leq \nu < 1$ ($\nu > 1$) [21].

It also has several nice statistical properties. First, it forms an exponential family. Second, it is more flexible than the Poisson because it handles both overdispersion and underdispersion. And third, the COM-Poisson has as special cases the following well-known distributions: the geometric, Poisson, and binomial correspond to $\nu = 0, 1$, and ∞, respectively. There are also several connections between the COM-Poisson and NB. First, the geometric is a special case of both; and of course, the sum of independent geometric variates is NB. Next, Imoto [10] proposed a three-parameter generalized COM-Poisson distribution in order to include the NB distribution as a special case of it.

2.1.1. *COM-Poisson regression with (α, ν) parametrization.* Guikema and Goffelt [9] developed a COM-Poisson generalized linear model (GLM) using a different parametrization, $(\alpha = \lambda^{1/\nu}, \nu)$. This new parametrization was introduced because α is good approximation of the mean when $\nu \leq 1$ or $\alpha > 10$ [21]. Thus, it acts as an approximate centering parameter [9]. They used a Bayesian approach for this GLM setting. It was later used to analyze motor vehicle crashes [11]. Barriga and Louzada [2] constructed a zero-inflated

COM-Poisson (ZICOM-Poisson) regression model and applied it to apple cultivar data using a Bayesian approach.

The pmf of the COM-Poisson based on (α, ν) is then (with slight abuse of notation)

$$P(X = k|\alpha, \nu) = Z(\alpha, \nu)^{-1} \left(\frac{\alpha^k}{k!}\right)^{\nu}, \quad \text{where} \quad Z(\alpha, \nu) = \sum_{k=0}^{\infty} \left(\frac{\alpha^k}{k!}\right)^{\nu}.$$

The general form of the COM-Poisson GLM is then [9]

$$\ln(\alpha) = \beta_0 + \sum_{i=1}^{p} \beta_i U_i \quad \text{and} \quad \ln(\nu) = \gamma_0 + \sum_{j=1}^{q} \gamma_j V_j \tag{2.1}$$

where U_i and V_j are covariates for $i = 1, \ldots, p$, and $j = 1, \ldots, q$: p (q) covariates are related to the centering (dispersion) link function. In some previous studies [9, 12, 17, 19], no covariates were assigned to the dispersion, in which case (2.1) reduces to

$$\ln(\alpha) = \beta_0 + \sum_{i=1}^{p} \beta_i U_i \quad \text{and} \quad \ln(\nu) = \gamma_0. \tag{2.2}$$

2.1.2. *COM-Poisson regression with (λ, ν) parametrization.* Sellers and Shmueli [19] used a COM-Poisson regression formulation based on the (λ, ν) parametrization by choosing a log-link function for both parameters. They extended the GLM formulation to the COM-Poisson case and indirectly modeled the relationship between the count X and the predictors U via $E(X)$. They also used maximum likelihood (ML) to obtain parameter estimates. Moreover, they discussed model estimation, inference, and diagnostics of this regression model, using several data sets as illustrations.

In this paper, we use this parametrization of the COM-Poisson distribution when fitting the COM-Poisson regression model to mRNA counts. We first assign covariates to both the centering and dispersion link functions as in (2.1), and then we compare its performance with no covariates assigned for the dispersion link function as in (2.2). The resulting likelihood function of the COM-Poisson GLM is not analytically tractable, so we use iterative numerical procedures to compute MLEs of the parameters and their standard errors. Computational details are given in Section 2.3.

We use a zero-inflated COM-Poisson (ZICOM-Poisson) regression model to handle excessive numbers of zeroes when they occur. We use a standard construction, so the pmf is:

$$P(X = k|\lambda, \nu, \zeta) = \begin{cases} \zeta + (1-\zeta) Z(\lambda, \nu)^{-1} & \text{if } k = 0 \\ (1-\zeta) Z(\lambda, \nu)^{-1} \left(\frac{\lambda^k}{k!}\right)^{\nu} & \text{if } k = 1, 2, \ldots, \end{cases} \tag{2.3}$$

where $\zeta \in [0, 1]$ models the excess-zero probability, and (λ, ν) are as before. For the GLM case, the parameters λ, ν, ζ could depend on the covariates. In this paper, we first assign a covariate to only (λ, ζ); we then assign a covariate to all three (λ, ν, ζ). The ZICOM-Poisson model with logistic link function and normal link function have also been studied [2] using a Bayesian approach. Recently, Sellers and Raim [18] studied the ZICOM-Poisson regression model also using ML.

2.2. The other models. We present the other well-known models here only to establish our notation. The NB is a gamma-Poisson mixture: if $X|\tau \sim \text{Poisson}(\tau)$ and τ has a gamma distribution with shape parameter $r > 0$ and scale parameter μ/r, then the unconditional pmf of X is

$$P(X = k|\mu, r) = \frac{\Gamma(k+r)}{\Gamma(k+1)\Gamma(r)} \left(\frac{r}{r+\mu}\right)^r \left(\frac{\mu}{r+\mu}\right)^k \quad \text{for} \quad k = 0, 1, 2, \ldots,$$

so that X has mean μ. For a fixed r, the NB likelihood can be expressed as a GLM form with a log link function: $\mu = \exp[U'\beta]$. The zero-inflated Poisson (ZIP) and NB (ZINB) can be constructed as in (2.3). Once again, we use the log link as we did for the COM-Poisson. (One referee suggested that we consider a two-link function for NB in order to make comparison of NB with COM-Poisson with two-link more fair. However, we found no previous work using a two-link NB GLM.)

Our last model is biophysically based. It arises as the steady-state mRNA density of coupled ordinary differential equations of certain chemical master equations for mRNA dynamics: see the supplement to [16]. The resulting pmf has the form

$$P(X = k|\theta, \mu) \propto \frac{1}{\mu\Gamma(\theta)} \left(\frac{k}{\mu}\right)^{\theta-1} e^{-k/\mu} \quad \text{for} \quad k = 0, 1, 2, \ldots, \tag{2.4}$$

where θ and μ are functions of the rate of gene activation and inactivation, the rate of mRNA transcription, and the rate of mRNA decay. This time, for fixed θ, we use the reciprocal link: $\mu = (U'\beta)^{-1}$.

2.3. Computation. We fit all models above using PROC COUNTREG procedure in SAS. Among several options, we chose the Newton-Raphson method to compute MLEs of the parameters and their standard errors. For the COM-Poisson and ZICOM-Poisson this procedure allows us to specify either the (λ, ν) or (α, ν) parametrization; in Section 3, we present summaries using the original (λ, ν) parametrization. We did encounter some computational difficulties with this procedure when fitting COM-Poisson regression because some parameter ranges led to slower run times. In addition, the procedure had difficulty in computing the Hessian matrix in a few cases. For NB and Poisson models, the running times were shorter than that of the COM-Poisson regression models. Once again, we encountered difficulty in computing the Hessian matrix in a few cases for these models.

We also used the R package to validate our computations. We used the *zeroinfl* function in the *pscl* package in R for the ZINB and ZIP. To fit COM-Poisson models, we used the *glm.comp* function in the *CompGLM* package in R. For ZICMP models, we used the same R code used in [18], which will be soon released by the authors.

3. mRNA count data analysis

In this section, we apply the models above to mRNA count data from So, et al. [24] and Raj, et al. [16]. We briefly describe the experimental conditions and the resulting mRNA data. We also provide some descriptive summary statistics to provide insight into these data sets. We then compare the performance of the fits of those models described in Section 2.

3.1. Comparing *E. coli* expression levels for different promoters. So, et al. [24] measured the expression levels of different promoters and under different conditions by conducting 20 experiments. They recorded the mRNA counts produced under each experimental condition. The aim of these experiments was to achieve different expression levels from the P_{lac} promoter. By using these expression levels, they compared mRNA lifetimes for the same transcript at different expression levels and growth rates. For illustration, we present mRNA counts of five of these experiments, in which different expression levels from different promoters are used to obtain the mRNA counts. In these five experiments, a bacterial strain, TK310, was grown with 0 to 1 millimolar (mM) of a reagent called IPTG or 0 to 10 mM of second messengers, which are intracellular signaling molecules called cAMP. More biophysical details about these experiments are in the supplementary material for [24].

In Experiment 1, the bacterial strain TK310, was grown with .1 mM of cAMP and seven different levels of IPTG: 0, 3, 10, 30, 100, 300, 1000 micromolar (μM). In Experiment 2, strain TK310 was grown with 1mM of IPTG and six different levels of cAMP: 0, 3, 10, 30, 100, 300 μM. Experiment 3 is similar to the Experiment 2, but allows higher levels of cAMP. Strain TK310 was grown with 1mM of IPTG and nine different levels of cAMP: 0, 3, 10, 30, 100, 300, 1000, 3000, 10000 μM. In Experiment 5, strain TK310 was grown with 1mM of IPTG and with seven different levels of cAMP: 0, 30, 100, 300, 1000, 3000, 10000 μM. Experiment 9 is similar to the Experiment 1; it uses 10 mM of cAMP with the same levels of IPTG as in Experiment 1. Basic descriptive statistics of these experiments are in Table 1, which clearly illustrates that the counts are considerably overdispersed in all experiments. Also, there are considerable numbers of zero mRNA counts. We will see below that there is rather strong evidence for excess zeroes.

Table 1. Summaries that suggest overdispersion and excessive zeroes.

Experiment	Mean	Variance	Zero counts	Sample size
1	1.50	11.66	4298	6092
2	12.99	232.55	1808	6492
3	13.00	338.42	4567	13464
5	11.96	307.44	2417	7852
9	19.40	771.02	3976	8960

To illustrate the usefulness of the COM-Poisson regression to modeling mRNA counts, we compare its performance to the Poisson and NB regression models. The COM-Poisson is first fit with no covariate assigned to the dispersion link (CMP1) then fit by assigning the same covariate to both link functions (CMP2). Later, we consider zero-inflated versions of the COM-Poisson, NB, and Poisson due to the excessive numbers of the zeroes in the data sets. For each data set, we compare these fits by assigning the corresponding covariate for each experiment. For Experiments 1 and 9, IPTG is the covariate because the cAMP is fixed. For Experiments 2, 3, and 5, cAMP is the covariate because IPTG is fixed. We note that, in the future, it would be better to do experiments in which both the cAMP and IPTG values are varied systematically so that a response-surface approach could examine both main effects and possible interactions.

The effects of the covariates are not of interest in this paper because our main purpose is model comparison. We first compare the model fits using the BIC, the values of which are in Table 2. For CMP2, we assign the same covariates to both link functions. It is clear that assigning a covariate to the dispersion link function improves the fit of the COM-Poisson model considerably. In all but one case, the COM-Poisson GLM with a covariate assigned to both link functions performs better than the NB. Further, we see that the COM-Poisson models and NB outperform the Poisson model by a wide margin.

Table 2. BIC of each model fit of COM-Poisson with no covariates assigned to dispersion link function (CMP1), COM-Poisson with covariates assigned to both link functions (CMP2), Negative Binomial (NB), and Poisson.

	CMP1	CMP2	NB	Poisson
experiment 1	13064	12131	11853	15061
experiment 2	40833	37609	38060	63169
experiment 3	72265	65885	66812	103424
experiment 5	42504	38736	38869	63182
experiment 9	57424	47442	47770	113196

Next, we used Vuong's closeness test [26] to support our conclusion made based on the BIC goodness-of-fit measure. This likelihood-ratio-based test for model selection uses an estimate of the Kullback-Leibler divergence. In all of these experiments, Vuong's test suggests that both COM-Poisson and the NB models are significant improvements over the Poisson model ($p < 0.001$ for all comparisons).

Table 3. BIC of each model fit of zero-inflated COM-Poisson with no covariates assigned to dispersion link function (ZICMP1), zero-inflated COM-Poisson with covariates assigned to both link functions (ZICMP2), zero-inflated Negative Binomial (ZINB), and zero-inflated Poisson (ZIP) models, respectively.

	ZICMP1	ZICMP2	ZINB	ZIP
experiment 1	11318	11323	11268	12967
experiment 2	36038	35877	35651	53470
experiment 3	67274	65467	66158	97014
experiment 5	40081	38610	38887	57106
experiment 9	57443	47448	45101	89044

Because of the excessive number of zeroes, we next consider zero-inflated versions of the NB (ZINB) and COM-Poisson (ZICMP). BICs for these models are presented in Table 3. It is clear that the models which account for zero-inflation are better than those without it. Among the zero-inflated models, the COM-Poisson with covariates assigned to both link functions is the best for all but one case. Note also that the use of covariates for both links in both the ZICMP and COM-Poisson models is substantially better than not using covariates for the dispersion link. Vuong's closeness test shows that ZICOM-Poisson models and ZINB model are highly significant improvements over the ZIP model ($p < 0.001$ for all comparisons).

3.1.1. *Summary of the model fits and diagnostics for Experiment 1.* For illustrative purposes, we present further details for Experiment 1. These include parameter estimates and their standard errors for the models. Also, we show some residual diagnostic plots. Similar results hold for the other experiments, so we omit them.

Fits of models without excess zeroes: In Table 4, the estimated model parameters and their standard errors are provided for each model. In this table, all the parameter estimates are highly significant. The slope estimates for IPTG are all positive, so increasing the IPTG level increases the mRNA counts. The Poisson and CMP1 models

appear to underestimate the uncertainty in the parameter estimates relative to the NB and CMP2. Following a referee's suggestion, we also fit the two COM-Poisson models using a Bayesian approach: the parameter estimates were very close to the MLEs presented here, so we omit them; this is not surprising because of the rather large sample size.

Table 4. Estimated model parameters and their standard errors given in parenthesis of COM-Poisson with no covariates assigned to dispersion link function (CMP1), COM-Poisson with covariates assigned to both link functions (CMP2), Negative Binomial (NB), and Poisson models, respectively.

	CMP1	CMP2	NB	Poisson
For mean link				
Intercept	-3.47 (.071)	-5.18 (.816)	-7.23 (.121)	-5.66 (.073)
IPTG	.48 (.063)	.74 (.098)	1.21 (.178)	0.97 (.010)
For dispersion link				
Intercept		-13.27 (.102)		
IPTG		1.54 (.016)		

Estimates of the dispersion parameters are of interest to us because they provide useful information about the direction of the dispersion. For CMP1, $\hat{\nu} = .18$; for CMP2, $\hat{\nu}$ ranges from .00 to .39; and for NB, $\hat{r} = 1.16$. Thus, mRNA count data in experiment 1 is strongly overdispersed, as is also clear from Table 2. This conclusion is also supported by using the likelihood ratio test (LRT) of $H_0 : \nu = 1$ *vs.* $H_1 : \nu \neq 1$. For CMP1 $-2\log\Lambda = 2006$ and the p-value < 0.001), so it is not equidispersed. The results for CMP2 is similar.

The Poisson model performs poorly with respect to the zero counts. Its estimate of the proportion of zero mRNA counts is 0.61, which is well below the observed proportion (0.71). The NB more closely estimates the proportion of zeros (0.69). The CMP1 and CMP2 also estimates the proportion of zeros (0.64 and 0.67, respectively) better than Poisson model, but they still underestimate the observed proportion. Among all the models considered in this section, we can see that Poisson model explains the data very poorly.

Fits of zero-inflated models: The estimated parameters and their standard errors are provided for each zero-inflated model in Table 5. All the parameter estimates in this table are highly significant. In all of these zero-inflated models, IPTG is always a statistically significant variable, and provides a positive slope. On the other hand, the IPTG effect on the zero component implies that zero mRNA counts are less likely as IPTG increases.

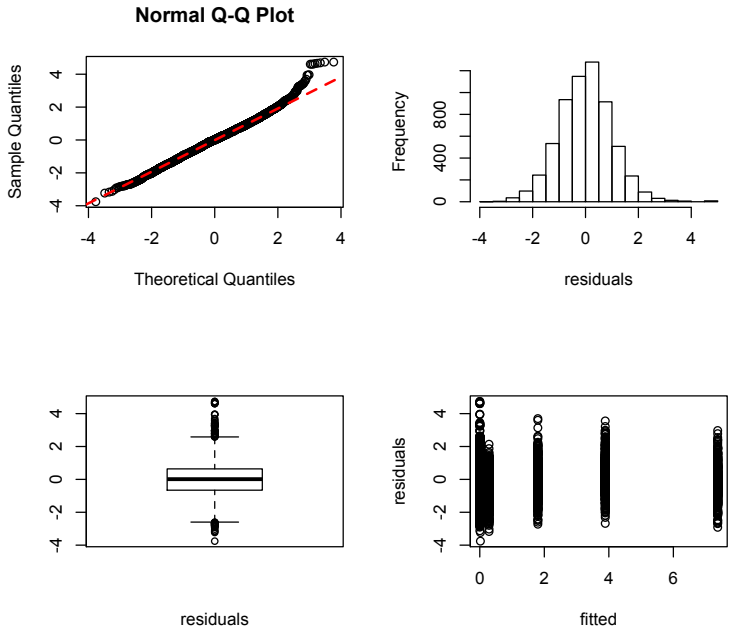

Figure 1. Diagnostic plots for the ZICMP1 model

Table 5. Estimated model parameters (and standard errors) of zero-inflated COM-Poisson with no covariates assigned to dispersion link function (ZICMP1), zero-inflated COM-Poisson with covariates assigned to both link functions (ZICMP2), zero-inflated Negative Binomial (ZINB), and zero-inflated Poisson (ZIP), respectively.

	ZICMP1	ZICMP2	ZINB	ZIP
Count component				
For mean link				
Intercept	-1.15 (.062)	-1.12 (.0.562)	-2.34 (.181)	-1.85 (.104)
IPTG	.17 (.009)	.15 (.006)	.54 (.025)	.48 (.014)
For dispersion link				
Intercept		-1.45 (.247)		
IPTG		0.06 (.032)		
Zero component				
Intercept	14.34 (.677)	14.20 (.597)	13.64 (.589)	11.43 (0.375)
IPTG	-2.55 (.125)	-2.53 (.107)	-2.42 (.104)	-1.95 (.063)

We next present several diagnostic plots for ZICMP1 model. In Figure 1, we see that the Q-Q plot of the Pearson residuals indicate an excellent fit over a wide range of quantiles. The histogram and boxplots of the residuals appear to be symmetric around zero, and there is no relationship between the residuals and fitted values. In short, we conclude that the ZIMCP1 model fits the data quite well.

Since $\hat{\nu} = .15$ in ZICMP1, $\hat{\nu}$ ranges from .25 to .38 in ZICMP2, and $\hat{r} = 2.06$ in NB, mRNA count data in experiment 1 is overdispersed. The LRT to test ZICMP model against the ZIP model ($H_0 : \nu = 1$ *vs.* $H_1 : \nu \neq 1$) was also highly significant (p-value < 0.001) for each of these data sets; so we conclude that data sets are not equidispersed. We also used the LRT to compare both ZICMP2 and ZINB with ZIP for each experiment. Once again, the LRT for these comparisons are all highly statistically significant, so we do not present the details.

In Experiment 1, all the models used in this section (ZIP, ZINB, ZICMP1, and ZICMP2) estimate the proportion of zero mRNA counts as 0.71 which is the observed proportion. Similar conclusions were reached for the other experiments. Thus, we see that the zero-inflated models estimates the proportion of zeros much closer to the observed proportion than NB, Poisson, and COM-Poisson models.

3.1.2. *Summary of all 20 experiments.* The results of the 20 experiments are in general consistent with the results of Experiment 1 summarized above. For COM-Poisson models, the estimates of ν are all between 0 and 1, indicating overdispersion in all data sets. We have the same conclusion with the NB, which is not surprising because that model only allows for overdispersion. The zero-inflated models yield estimated proportion of zeroes that are closer to the observed proportions. Almost all estimates of the regression model parameters are highly significant. We also did similar model comparisons for the experiments not included in Tables 2 or 3. Below is a brief summary of our findings.

- In 18 experiments, the COM-Poisson with a covariate assigned to both link functions performs considerably better than the COM-Poisson with no covariate assigned to dispersion link. A similar conclusion holds for the zero-inflated case.
- In 9 experiments, the COM-Poisson with a covariate assigned to both link functions performs better than the NB; in 2 experiments, their performance is similar, with the difference in BIC is less than 10. Here again, the same conclusion holds for the zero-inflated cases of these models.
- In the majority of the experiments, the performance of COM-Poisson with with no covariate assigned to dispersion link and its zero-inflated version are much poorer than the NB and ZINB, respectively.

The Poisson model performs poorly in all experiments. Also, the COM-Poisson with a covariate assigned to both link functions performs similar to the NB. This is plausible because COM-Poisson with small values of ν are close to a geometric distribution, a special case of the NB.

3.2. Comparing different doxycycline levels in mammalian cells. Cells from a homogeneous population can express different numbers of molecules of specific proteins. Raj et al. [16] has studied these variations by counting individual molecules of mRNA produced from a reporter gene. They considered two cell lines: E-YFP-M1-1x and E-YFP-M1-7x, which we call gene line 1 and gene line 7, respectively. They found that the variability across the population remained constant for all doxycycline concentration levels for gene line 1, but that it varied non-monotonically for gene line 7.

In their experiments, they varied the doxycycline concentration levels in units of nanograms per milliliter (ng/ml) thus: 0, 0.02, 0.04, 0.08, 0.16, 0.32. Then they measured the number of mRNA molecules per cell. There are 614 records collected from gene-line 1 with mRNA counts ranging from 0 to 313; and there are 608 records collected from gene-line 7 with counts from 0 to 1031. From detailed biophysical considerations, they derived a model for the count distribution, which they then approximated by a pmf that resembles a gamma density (2.4). Here, we compare the GLM associated with this

biophysically derived pmf with the others, as in Table 2. The fits are summarized in Table 6.

Table 6. BIC of each model fit of COM-Poisson with no covariates assigned to dispersion link function (CMP1), COM-Poisson with covariates assigned to both link functions (CMP2), Negative Binomial (NB), biophysical model (2.4), and Poisson for mammalian cells.

	CMP1	CMP2	NB	(2.4)	Poisson
gene line 1	5287	5284	5292	5197	16271
gene line 7	5386	5331	5006	4560	52170
combined	10680	10609	10554	9947	69561

From Table 6, it is clear that the fit of COM-Poisson improves when we introduce in the dispersion link function; however, neither COM-Poisson or NB is as good as the discrete gamma fit (2.4). Among all these models, the Poisson performs very poorly. We also ran the regression models with two covariates: one for the the doxycycline levels and the other for the two gene lines. The fits are summarized in the last row of Table 6. Once again, the biophysical model has the best fit among these models.

We did find that the mRNA counts are significantly different for doxycycline concentration levels. However, we omit the the parameter estimates. We include this example primarily to emphasize the importance of using a model that is derived from biophysical considerations, which in this case outperforms the other models commonly used for count data.

4. Discussion

Modeling biological phenomena can be complicated because there are many factors that affect outcomes and that are hard to control. In the experimental data that we study here, the variation between genes which are from the same population makes it unlikely that a single model is best universally. Our aims in this paper are to introduce the COM-Poisson distribution, to model zero inflation, and to consider regression methods on mRNA count data under different experimental conditions. In particular, we show that the use of covariates in both link functions for the COM-Poisson or ZICMP GLM is much better than assigning no covariate to the dispersion link. We compare these regression models with the more commonly used Poisson, ZIP, NB and ZINB GLM. For *E. coli* data, we see that the ZICMP GLM is as good as the ZINB GLM. COM-Poisson models perform much better than Poisson and as good as NB models.

In the absence of detailed biophysical knowledge, the COM-Poisson regression model is often a good candidate for fitting over dispersed mRNA count data. And when we do know more about the biophysics (as for the mammalian cell data), this analysis can help to confirm the adequacy of the approximations to probability distributions derived from master equations. The COM-Poisson model was first proposed in the queueing theory context [4]. Queueing models have also been considered for gene expression and mRNA transcription [5]. Given the good fits of the COM-Poisson regression model and its variants above, we suggest that it is worthwhile to pursue this connection with queueing theory to describe mRNA dynamics. One possible direction is to do a closer study of the biophysical models to possibly use biophysical considerations to propose other Markov processes that are variants of the telegraph model that may lead to better fits to mRNA counts.

Appendix: A model for burstiness that leads to overdispersion.

There is a relatively simple model that connects burstiness with overdispersion. As stated in the introduction, the two prominent models proposed for mRNA transcription are Markovian: the Poisson and the telegraph processes [5, 8, 27]. The telegraph model is a two-state Markov process with the two states representing the active and inactive phases of the promoter.

For $t \geq 0$, consider a two-state Markov process $\lambda(t)$ with two states, $0 \leq a < b$. Here, a (b) corresponds to the off-state (on-state) when there is very little or no (large) mRNA production. Next, let the counting process $\{N_t : t \geq 0\}$ be a doubly stochastic Poisson process with random intensity $\lambda(t)$. For any finite $T > 0$, let

$$\Lambda(T) = \int_0^T \lambda(t)\, dt$$

be the cumulative intensity. Then, the first two moments of N_t are

$$E[N_T] = EE[N_T|\Lambda(T)] = E[\Lambda(T)]$$

and

$$\text{var}[N_T] = E(\text{var}[N_T|\Lambda(T)]) + \text{var}(E[N_T|\Lambda(T)]) = E[\Lambda(T)] + \text{var}[\Lambda(T)].$$

Because $\text{var}[\Lambda(T)] > 0$ the variance of the count is larger than the mean of the count, and we have overdispersed counts. Moreover, the Fano factor (the ratio of the variance to the mean) is greater than 1. For an example of a more detailed biophysical model that leads to Fano factor greater than 1 see [13].

Acknowledgment We extracted data from So, et al. [24] and from Raj, et al. [16]. We thank Professor Ido Golding for sending us data from his laboratory and for his advice; and Professor Hanna Salman for informative conversations. We also thank the referees for their helpful comments, which improved our paper considerably. Dr. Simsek was supported by an Andrew Mellon Predoctoral Fellowship at the University of Pittsburgh.

References

[1] Arazia, A., Ben-Jacob, E.B. and Yechiali, U. *Bridging genetic networks and queueing theory*, Physica A 332, 585-616, 2004.

[2] Barriga, G.D.C. and Louzada, F. *The zero-inflated Conway-Maxwell-Poisson distribution: Bayesian inference, regression modeling and influence diagnostic*, Statistical Methodology **21**, 23-34, 2014.

[3] Chatla, S.B. and Shmueli, G. *An efficient estimation of Conway-Maxwell-Poisson regession and additive model with an application to bike sharing*, arXiv:1610.08244v2, 2016.

[4] Conway, R.W. and Maxwell, W.L. *A queuing model with state dependent service rates*, Journal of Industrial Engineering **12**, 132-136, 1962.

[5] Elgart, V., Jia, T., and Kulkarni, R.V. *Applications of Little's law to stochastic models of gene expression*, Physical Review E **82**, 021901(1-6), 2010.

[6] Fraser D. and Kærn M. *A chance at survival: gene expression noise and phenotypic diversification strategies*, Molecular Microbiology **71**, 1333-1340, 2009.

[7] Gaunt, R.E., Iyengar, S., Olde Daalhuis, A.B., and Simsek, B. *An asymptotic expansion for the normalizing constant of Conway-Maxwell-Poisson distribution*, arXiv:1612:06618, 2016.

[8] Golding, I., Paulsson, J., and Cox, E.C. *Real-time kinetics of gene activity in individual bacteria*, Cell **123**, 1025-1036, 2005.

[9] Guikema, S.D. and Goffelt, J.P. *A flexible count data regression model for risk analysis*, Risk Analysis **28**, 213-223, 2008.

[10] Imoto, K. *A generalized Conway-Maxwell-Poisson distribution which includes the negative binomial distribution*, Applied Mathematics and Computation **247**, 824-834, 2014.

[11] Lord, D., Guikema, S.D., and Geedipally, S.R. *Application of the Conway-Maxwell-Poisson generalized linear model for analyzing motor vehicle crashes*, Accident Analysis & Prevention **40**, 1123-1134, 2008.

[12] Minka, T., Shmueli, G., Kadane, J.B., Borle, S., and Boadwright, P. *Computing with the COM-Poisson distribution*, Technical Report 776, Statistics Department, Carnegie Mellon University, 2003.

[13] Mitarai N, Semsey A, Sneppen K. (2015) Dynamic competition between transcription initiation and repression: Role of non-equilibrium steps in cell to cell heterogeneity. arXiv: 1502.03011v3.

[14] Peccoud, J. and Ycart, B. *Markovian modeling of gene-product synthesis*, Theoretical Population Biology **48**, 222-234, 1995.

[15] Pogany T.K. *Integral form of the COM-Poisson renormalization constant*, Statistics & Probability Letters **119**, 144-145, 2016.

[16] Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y., and Tyagi, S. *Stochastic mRNA synthesis in mammalian cells*, PLOS Biology **4**, 1707-1719, 2006.

[17] Sellers, K.F., Borle, S., and Shmueli, G. *The COM-Poisson model for count data: a survey of methods and applications*, Applied Stochastic Models in Business and Industry **28**, 104-116, 2012.

[18] Sellers, K.F. and Raim, A.M. *A flexible zero-inflated model to address data dispersion*, Computational Statistics and Data Analysis **99**, 68-80, 2016.

[19] Sellers, K.F. and Shmueli, G. *A flexible regression model for count data*, Annals of Applied Statistics **4**, 943-961, 2010.

[20] Shahrezaei, V. and Swain, P.S. *Analytical distributions for stochastic gene expression*, Proceedings of the National Academy of Sciences **45**, 17256-17261, 2008.

[21] Shmueli, G., Minka, T., Kadane, J.B., Borle, S., and Boatwright, P. *A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution*, Journal of the Royal Statistical Society C **54**, 127-142, 2005.

[22] Simsek B. *Applications of Point Process Models to Imaging and Biology*, PhD Dissertation, Statistics Department, University of Pittsburgh, 2016.

[23] Simsek, B. and Iyengar, S. *Approximating the the Conway-Maxwell-Poisson normalizing constant*, Filomat **30**, 953-960, 2016.

[24] So, L., Ghosh, A., Zong, C., Sepulveda, L.A., Segev R., and Golding I. *General properties of transcriptional time series in Escherichia coli*, Nature Genetics **43**, 554-560, 2011.

[25] Trcek, T., Chao, J.A., Larson, D.R., Park, H.Y., Zenklusen, D., Shenoy, S.M., and Singer, R.H. *Single-mRNA counting using fluorescent in situ hybridization in budding yeast*, Nature Protocols **7**, 408-419, 2012.

[26] Vuong, Q.H. *Likelihood ratio tests for model selection and non-nested hypotheses*, Econometrica. **57**, 307-333, 1989.

[27] Zenklusen, D., Larson, D.R., and Singer, R.H. *Single-RNA counting reveals alternative modes of gene expression in yeast*, Natural Structural & Molecular Biology **15**, 1263–1271, 2008.

[28] Zhang, H., Pounds, S.B., and Tang, L. *Statistical methods for overdispersion in mRNA-Seq count data*, Open Bioinformatics Journal **7**, 34-40, 2013.