CHARACTERIZING THE PERFORMANCE OF THE BAYESIAN CONWAY-MAXWELL POISSON GENERALIZED LINEAR MODEL

Srinivas Reddy Geedipally[1]
Research Assistant
Zachry Department of Civil Engineering
Texas A&M University
3136 TAMU
College Station, TX 77843-3136
Tel. (979) 845-9892
Fax. (979) 845-6481
Email: srinivas8@tamu.edu

Seth D. Guikema
Assistant Professor
Department of Geography and Environmental Engineering
Johns Hopkins University
Baltimore, MD 21218
Tel. (410) 516-6042
Fax. (410) 516-8996
E-mail: sguikema@jhu.edu

Soma Sekhar Dhavala
Research Assistant
Department of Statistics
Texas A&M University
3143 TAMU
College Station, TX 77843-3143
Tel. (979) 845-3141
Fax. (979) 845-3144
Email: soma@stat.tamu.edu

Dominique Lord
Assistant Professor
Zachry Department of Civil Engineering
Texas A&M University
3136 TAMU
College Station, TX 77843-3136
Tel. (979) 458-3949
Fax. (979) 845-6481
Email : d-lord@tamu.edu

May 1, 2008

[1] Corresponding author

ABSTRACT

This paper documents the performance of a Bayesian Conway-Maxwell-Poisson (COM-Poisson) generalized linear model (GLM). This distribution was originally developed as an extension of the Poisson distribution in 1962 and has a unique characteristic, in that it can handle both under-dispersed and over-dispersed count data. Previous work by the authors lead to the development of a dual-link GLM based on the COM-Poisson distribution and applied this model to analyzing power system reliability and motor vehicle crash data. Parameter estimation for this model is done within the Bayesian framework using Markov Chain Monte Carlo methods. The objectives of this paper are to (1) characterize the parameter estimation accuracy of the Markov Chain Monte Carlo (MCMC) implementation of the COM GLM and (2) estimate the computational burden of this MCMC implementation. We use simulated datasets to assess the performance of the COM GLM. The results of the study indicate that the COM GLM is flexible enough to model under-, equi- and over-dispersed datasets with different sample mean values. The results also show that the MCMC implementation of the COM GLM yields accurate parameter estimates. Furthermore, we show that a previously suggested asymptotic approximation of the mean of the COM-Poisson distribution holds true even when the sample mean of the data is substantially below the lower bound previously suggested. However, the approximation is less accurate for very lower sample mean values, and we characterize the degree of this inaccuracy. Our results also show that the computational burden of the COM GLM is not prohibitive. The COM GLM provides a promising and flexible approach for performing count data regression.

Keywords: Generalized linear model, Conway-Maxwell-Poisson distribution, count data, Bayesian estimation, Markov Chain Monte Carlo, regression

1. INTRODUCTION

The Poisson family of discrete distributions stands as a benchmark for analyzing count data. The Conway-Maxwell Poisson (COM-Poisson) distribution is a generalization of the Poisson distribution and was originally developed in 1962 (Conway and Maxwell 1962) as a method for modeling both under-dispersed and over-dispersed count data. It was then "rediscovered" by Shmueli et al. (2005), where many of the properties of the distribution were also first derived. The COM-Poisson belongs to the exponential family as well as to the two-parameter power series family of distributions. This distribution introduces an extra parameter ν which governs the rate of decay of successive ratios of probabilities. It nests the usual Poisson ($\nu = 1$), geometric ($\nu = 0$) and Bernoulli ($\nu = \infty$) distributions. The COM-Poisson distribution allows for both thicker and thinner tails than those of the Poisson distribution (Boatwright et al., 2003; Shmueli et al., 2005). The conjugate priors for the parameters of the COM-Poisson distribution have also been derived (Kadane et al., 2006). The COM-Poisson distribution has been used in many studies such as analyzing word length (Shmueli et al., 2005), birth process models (Ridout and Besbeas, 2004), prediction of purchase timing and quantity decisions (Boatwright et al., 2003), quarterly sales of clothing (Shmueli et al., 2005), internet search engine visits (Telang et al., 2004), the timing of bid placement and extent of multiple bidding (Borle et al., 2006), modeling electric power system reliability (Guikema and Coffelt 2008) and modeling motor vehicle crashes (Lord et al., 2008). Only Guikema and Coffelt (2008) and Lord et al. (2008) have used the COM-Poisson in a regression setting.

Guikema and Coffelt (2008) introduced a generalized linear model (GLM) built on the COM-Poisson, and Lord et al. (2008) then utilized this model to analyze traffic accident data. The COM GLM of Guikema and Coffelt (2008) is a full Bayesian model implemented in WinBUGS (Spiegelhalter et al., 2003). This paper evaluates the estimation accuracy and computational burden of the COM GLM for datasets characterized by over-, under- and equi-dispersion with different means. It also characterizes the accuracy of the asymptotic approximation of the mean of the COM-Poisson suggested by Shmueli et al. (2005).

This paper is organized as follows. The next section describes the COM-Poisson distribution and its GLM framework. The third section presents our research methodology. The fourth section gives the results of our computational study. The fifth section gives a brief discussion of the results, and the sixth section provides concluding comments.

2. BACKGROUND

This section describes the characteristics of the COM-Poisson distribution and the COM GLM framework.

2.1 PARAMETERIZATION

The COM-Poisson distribution was first introduced by Conway and Maxwell (1962) for modeling queues and service rates. Although the COM-Poisson distribution is not particularly new, it was largely unstudied and unused until Shmueli et al. (2005) derived the basic properties of the distribution. Kadane et al. (2006) then developed the conjugate priors for the parameters of the COM-Poisson distribution. Our description of the COM-Poisson that follows builds from the work of Shmueli et al. (2005).

The COM-Poisson distribution is a two-parameter extension of Poisson distribution that generalizes some well-known distributions including the Poisson, Bernoulli, and geometric distributions (Shmueli et al., 2005). It also offers a more flexible alternative to distributions derived from these discrete distributions (i.e. the binomial and negative binomial distributions). The COM-Poisson distribution can handle both under-dispersion (variance less than the mean) and over-dispersion (variance greater than the mean). The probability mass function (PMF) of the COM-Poisson for the discrete count Y is given by Equations (1) and (2)

$$P\left(Y=y\right)=\frac{1}{Z\left(\lambda,\nu\right)}\frac{\lambda^{y}}{\left(y!\right)^{\nu}} \tag{1}$$

$$Z\left(\lambda,\nu\right)=\sum_{n=0}^{\infty}\frac{\lambda^{n}}{\left(n!\right)^{\nu}} \tag{2}$$

where λ is a centering parameter that is related directly to the mean of the observations and ν is the shape parameter of the COM-Poisson distribution. The condition ν >1 corresponds to under-dispersed data, ν <1 to over-dispersed data, and ν =1 to equi-dispersed (Poisson) data. Several common PMFs are special cases of the COM-Poisson with the original formulation. Specifically, setting $\nu = 0$ yields the geometric distribution, $\lambda < 1$ and $\nu \rightarrow \infty$ yields the Bernoulli distribution in the limit, and $\nu = 1$ yields the Poisson distribution. This flexibility greatly expands the types of problems for which the COM-Poisson distribution can be used to model count data.

The asymptotic expressions for the mean and variance of the COM-Poisson derived by Shmueli et al. (2005) are given by Equations (3) and (4) below.

$$E\left[Y\right]=\frac{\partial\log Z}{\partial\log\lambda} \tag{3}$$

$$Var\left[Y\right]=\frac{\partial^{2}\log Z}{\partial\log^{2}\lambda} \tag{4}$$

The COM-Poisson distribution does not have closed-form expressions for its moments in terms of the parameters λ and ν. However, the mean can be approximated through a few different approaches, including (i) using the mode, (ii) including only the first few terms of Z when ν is large, (iii) bounding E[Y] when ν is small, and (iv) using an asymptotic expression for Z in Equation (1). Shmueli et al. (2005) used the last approach to derive the approximation in Equation (5).

$$E\left[Y\right] \approx \lambda^{1/\nu} + \frac{1}{2\nu} - \frac{1}{2} \tag{5}$$

Using the same approximation for Z as in Shmueli et al. (2005), the variance can be approximated as:

$$Var\left[Y\right] \approx \frac{1}{\nu}\lambda^{1/\nu} \tag{6}$$

These approximations may not be accurate for ν>1 or $\lambda^{1/\nu} < 10$ (Shmueli et al. 2005).

Despite its flexibility and attractiveness, the COM-Poisson has limitations in its usefulness as a basis for a GLM, as documented in Guikema and Coffelt (2008). In particular, neither *λ* nor *ν* provide a clear centering parameter. While *λ* is approximately the mean when *ν* is close to one, it differs substantially from the mean for small *ν*. Given that *ν* would be expected to be small for over-dispersed data, this would make a COM-Poisson model based on the original COM-Poisson formulation difficult to interpret and use for over-dispersed data.

Guikema and Coffelt (2008) proposed a re-parameterization using a new parameter μ = $\lambda^{1/\nu}$ to provide a clear centering parameter. This new formulation of the COM-Poisson is summarized in Equations (7) and (8) below:

$$P\left(Y = y\right) = \frac{1}{S\left(\mu, \nu\right)}\left(\frac{\mu^{y}}{y!}\right)^{\nu} \tag{7}$$

$$S\left(\mu, \nu\right) = \sum_{n=0}^{\infty}\left(\frac{\mu^{n}}{n!}\right)^{\nu} \tag{8}$$

By substituting $\mu = \lambda^{1/\nu}$ in equations (4), (5), and (41) of Shmueli (2005), the mean and variance of *Y* are given in terms of the new formulation as $E\left[Y\right] = \frac{1}{\nu}\frac{\partial \log S}{\partial \log \mu}$ and $V\left[Y\right] = \frac{1}{\nu^{2}}\frac{\partial^{2} \log S}{\partial \log^{2} \mu}$ with asymptotic approximations $E\left[Y\right] \approx \mu + 1/\left(2\nu\right) - 1/2$ and $Var\left[Y\right] \approx \mu/\nu$ especially accurate once *μ>10*. With this new parameterization, the integral part of μ is the mode leaving μ as a reasonable centering parameter. The substitution $\mu = \lambda^{1/\nu}$ also allows *ν* to keep its role as a shape parameter. That is, if ν < 1, the variance is greater than the mean while ν > 1 leads to under-dispersion. In this paper we investigate the accuracy of the approximation $E\left[Y\right] \approx \mu + 1/\left(2\nu\right) - 1/2$.

This new formulation provides a good basis for developing a COM GLM. The clear centering parameter provides a basis on which the centering link function can be built, allowing ease of interpretation across a wide range of values of the shape parameter. Furthermore, the shape parameter *v* provides a basis for using a second link function to allow the amount of over-dispersion, equi-dispersion or under-dispersion to vary across measurements.

2.2 GENERALIZED LINEAR MODEL

Guikema and Coffelt (2008) developed a COM GLM framework for modeling discrete count data using the reformulation of the COM-Poisson given in equations (7) and (8). This dual-link GLM framework, in which both the mean and the variance depend on covariates, is given in equations (9-12), where *Y* is the count random variable being modeled and x_i and z_j are covariates. There are *p* covariates used in the centering link function and *q* covariates used in the shape link function. The sets of parameters used in the two link functions do not need to be identical. If a single-link model is desired, the second link given by equation (12) can be removed allowing a single *v* to be estimated directly.

$$P(Y=y)=\frac{1}{S(\mu,\nu)}\left(\frac{\mu^y}{y!}\right)^{\nu} \tag{9}$$

$$S(\mu,\nu)=\sum_{n=0}^{\infty}\left(\frac{\mu^n}{n!}\right)^{\nu} \tag{10}$$

$$\ln(\mu)=\beta_0+\sum_{i=1}^{p}\beta_i x_i \tag{11}$$

$$\ln(\nu)=\alpha_0+\sum_{j=1}^{q}\alpha_j z_j \tag{12}$$

The GLM described above is highly flexible and readily interpreted. It can model under-dispersed datasets, over-dispersed datasets, and datasets that contain intermingled under-dispersed and over-dispersed counts (for dual-link COM-Poisson models only). The variance is allowed to depend on the covariate values, which can be important if high (or low) values of some covariates tend to be variance-decreasing while high (or low) values of other covariates tend to be variance-increasing. The parameters have a direct link to either the mean or the variance, providing insight into the behavior and driving factors in the problem, and the mean and variance of the predicted counts are readily approximated based on the covariate values and regression parameter estimates.

Parameter estimation in the COM GLM presented above is challenging. The likelihood equation for the COM GLM is complex, making analytical and numerical maximum likelihood estimation difficult. Thus, Bayesian estimation provides an attractive alternative for estimating the coefficients of the model. Guikema and Coffelt (2008) implemented the COM GLM in WinBUGS using a custom-coded COM-Poisson distribution. This paper characterizes the estimation accuracy and computational burden of the Markov Chain Monte Carlo (MCMC) COM GLM of Guikema and Coffelt (2008). It also investigates the degree of inaccuracy in using the asymptotic mean approximation developed by Shmueli et al. (2005). The next section describes the methodology used in this study. Despite its complex derivation, Sellers and Shmueli (2008) are currently developing a MLE approach for parameter estimation for the COM-Poisson distribution.

3. METHODOLOGY

To test the estimation accuracy and computational burden of the MCMC implementation of the COM GLM of Guikema and Coffelt (2008), we simulated a number of datasets from the COM GLM with known regression parameters that correspond to a wide range of mean and variance values. We then estimated the regression parameters of the COM GLM using the MCMC implementation. The estimated parameters were then compared to the known parameter values, and the computational burden of the MCMC was assessed in all the cases. In this section, we give details of the various steps involved.

3.1 DATA SIMULATION

In order to characterize the accuracy of the parameter estimates from the COM GLM, five different datasets were randomly generated for each of nine different scenarios. The nine scenarios include simulated datasets of under-dispersed, equi-dispersed and over-dispersed data. For each level of dispersion, three different sample means were used: high mean (~ 20.0), moderate mean (~ 5.0) and low mean (~ 0.8). Due to the high computational time and lack of readily available software, we restricted our analysis to five simulation runs (or datasets) for each scenario. Each of these five datasets was then used as input for the COM GLM, and the resulting parameters estimates were compared to the known parameters values that had be used to generate the datasets.

We simulated 1,000 values of the covariates X_1 and X_2 from a uniform distribution on [0, 1]. The centering parameter μ and shape parameter v were then generated according to Equations (11) and (12) with known (assigned) regression parameters. Note that the same covariates are used for both the centering and shape parameters. Realizations from the COM-Poisson are then generated using the inverse CDF method (Devroye, 1986).

The regression parameter values were selected in such a way that the shape parameter v was always set between 0 and 1 for simulating the over-dispersed datasets, above 1 for the under-dispersed datasets and approximately 1 for the equi-dispersed datasets. The parameters that were assigned in simulating the datasets are given in the table below. Note that we have assumed a single-link model by assigning values of zero to α_1 and α_2, respectively. However, we left these parameters in the MCMC COM GLM in order to test both (1) the computational burden of the full dual-link model and (2) the ability of the COM-Poisson model to accurately estimate zero values for these two parameters. Table 1 summarizes the characteristics of the simulation scenarios.

Table 1: Assigned parameters of the simulated datasets

	Over-dispersed data			Under-dispersed data			Equi-dispersed data		
	High mean	Moderate mean	Low mean	High mean	Moderate mean	Low mean	High mean	Moderate mean	Low mean
β_0	3.0	1.3	-2.0	3.0	1.7	0.2	3.0	1.7	0.2
β_1	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15
β_2	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15
α_0	-0.4	-1.3	-1.3	1.0	1.0	1.2	0	0	0
α_1	0	0	0	0	0	0	0	0	0
α_2	0	0	0	0	0	0	0	0	0

All the simulations were carried in MATLAB® 7.1.0 R14 (The Mathworks Inc , Natick, MA)

3.2 TESTING PROTOCOL

The MCMC implementation of the COM GLM proposed by Guikema and Coffelt (2008) was used for the model estimation process. The coefficients of the COM GLM were estimated using WinBUGS (Spiegelhalter et al., 2003). Non-informative priors (i.e., N(0,100) priors) were utilized for the parameters of COM GLMs. A total of 3 Markov chains were used in the model estimation process with 50,000 iterations per chain and no thinning. The first 25,000 iterations (burn-in samples) were discarded. The remaining 25,000 iterations were used for estimating the coefficients. The Gelman-Rubin (G-R) convergence statistic was used to verify that the simulation runs converged properly.

4. RESULTS

This section first describes the selection of the error term used in approximating the normalizing term (S) in the COM-Poisson distribution (see Equation (8)). This is followed by an assessment of the performance of COM-Poisson for the nine scenarios mentioned above. The results concerning the accuracy of the asymptotic approximation of the mean of the COM-Poisson suggested by Shmueli et al. (2005) and the computational burden of COM-Poisson in WinBUGS are then discussed.

4.1 ERROR TERM

The S term in the COM-Poisson distribution is the sum of an infinite series. However, the contributions of new terms in the series decreases as more terms are added the series. In order to approximate S, we used an iterative approach in which the change in S was monitored as new terms were added. The series was truncated when the contribution of new terms dropped below a predefined threshold, given as a fraction of the previous series value. We refer to this threshold as the relative error ε. Four levels of error were considered in this study: ε= 0.1, 0.01, 0.001, and 0.0001. The first set of runs of the COM-Poisson MCMC was performed to determine the effect of ε on both the computational burden of the process and the parameter estimates. These runs were accomplished using an over-dispersed high mean dataset on a computer with a 1.50 GHz Pentium 4 CPU and 512 MB of RAM. As seen in Figure 1, the error term does not have much effect on the parameter estimation accuracy. The estimated parameters were almost the same at all error levels.

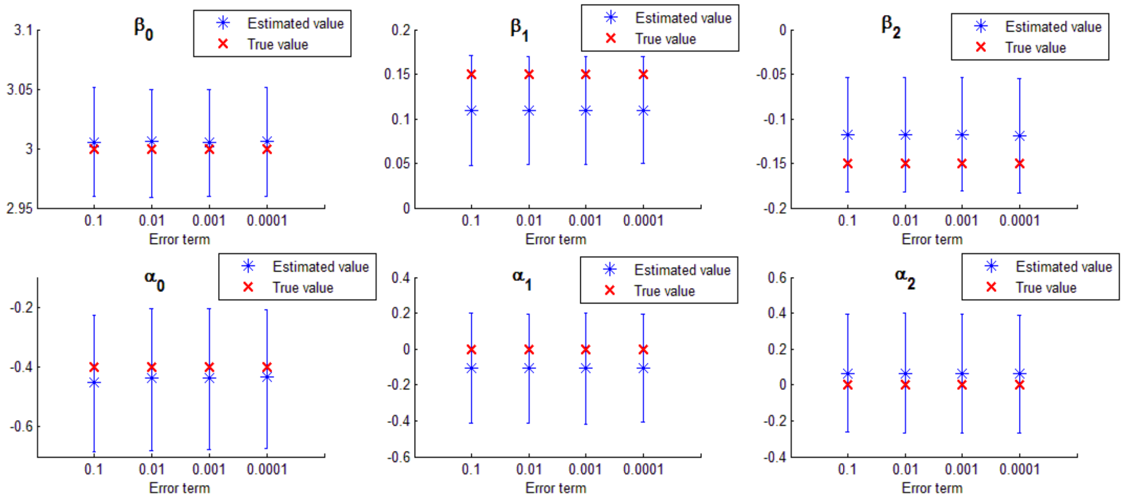


Figure 1: Parameter estimation accuracy versus relative error

The computational time was also not much different from one error level to another as shown in Figure 2. Note that the computational time depends on the computer system configuration on which the simulations were performed. We chose an error term of 0.01 in our analysis, although choosing a different error term value would not affect the results substantially.

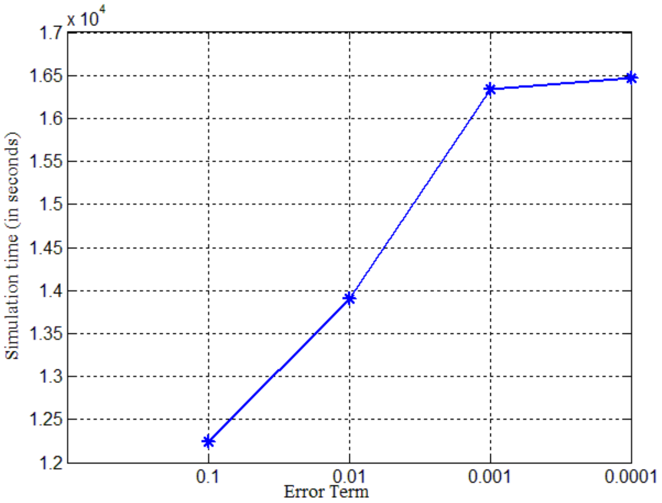


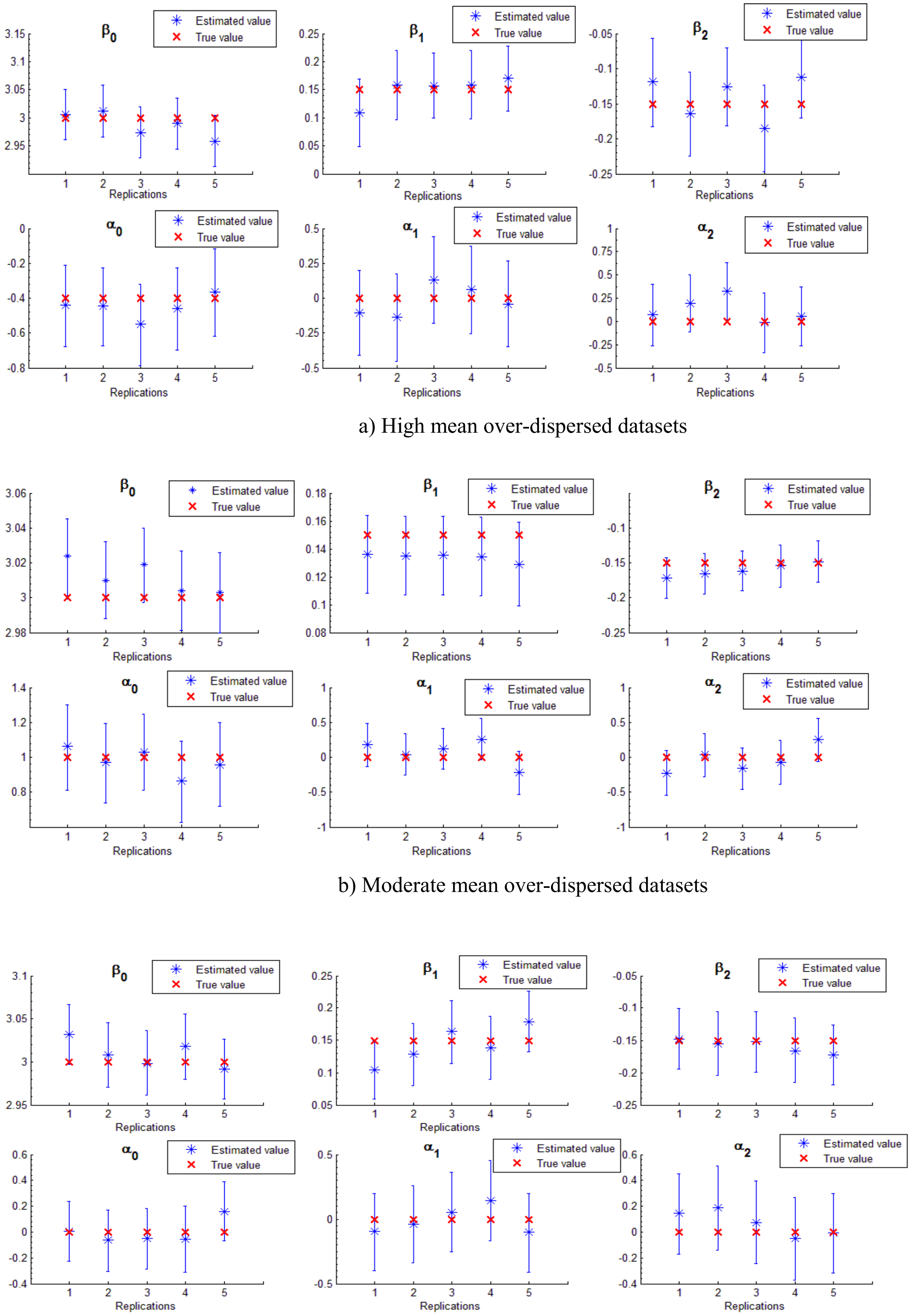
Figure 2: Simulation time for different error terms of COM-Poisson distribution

4.2 PARAMETER ESTIMATION ACCURACY

The estimated parameters and their 95% credible intervals were plotted and compared with the true parameters as shown in Figures 3-5. Each figure corresponds to a specific dispersion present in the data and each subplot corresponds to different sample mean value of the dataset. Figure 3 shows the plot for the over-dispersed datasets. It shows that the true parameters lie in the 95% credible interval for all cases and are generally close to the estimated posterior mean of the parameters. The credible intervals were found to be wider at low mean values for both the centering and shape parameter coefficients

Figure 4 below gives the plots for the under-dispersed datasets. Similar to the result above, the true parameters lie in the 95% credible interval for all cases and are generally close to the estimated posterior mean of the parameters. The credible intervals of the parameters were found to be wider (as expected) for low mean values for both centering and shape parameters.

Figure 5 shows the similar characteristics for the equi-dispersed datasets as that of other datasets. Except in one or two cases, all plots show that the true parameter lies inside the 95% credible intervals of estimated parameters. Although the problem with these exceptional cases was unknown yet, it could be attributed to the randomness in the datasets. Also, the credible intervals of the parameters were found to be wider for the low mean values for both centering and shape parameters.

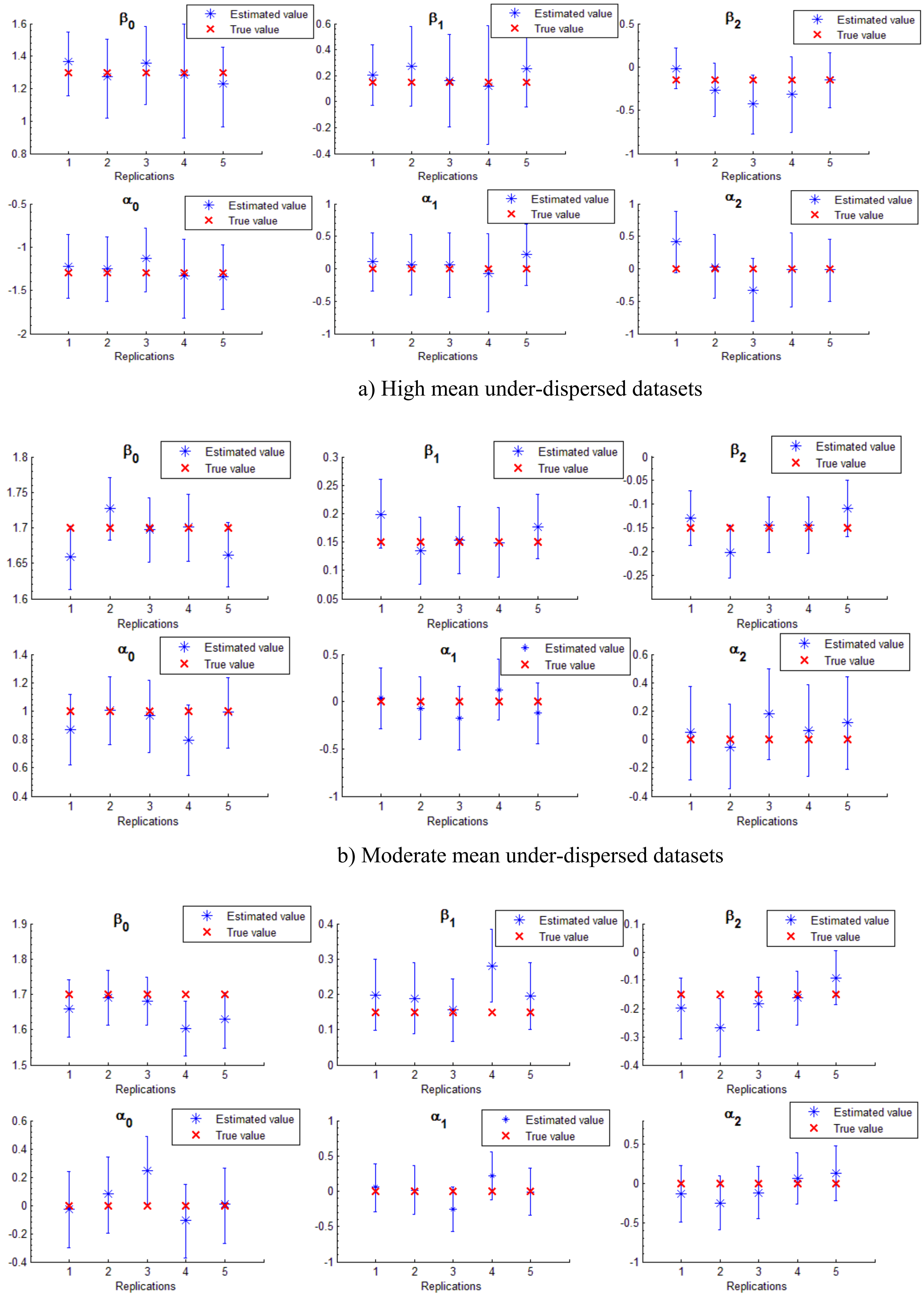


a) High mean over-dispersed datasets

b) Moderate mean over-dispersed datasets

c) Low mean over-dispersed datasets

Figure 3: Parameters estimates of over-dispersed datasets

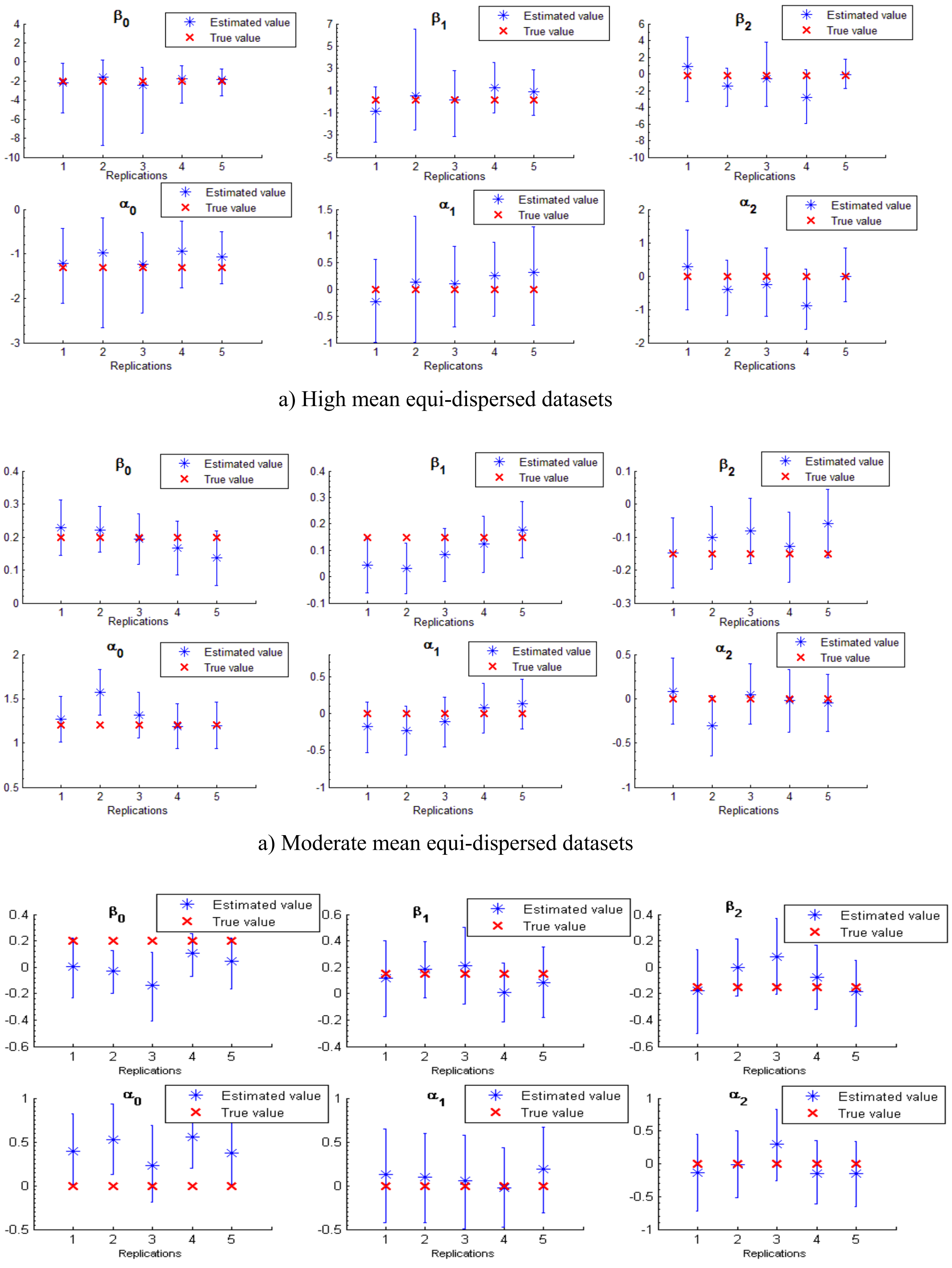


a) High mean under-dispersed datasets

b) Moderate mean under-dispersed datasets

c) Low mean under-dispersed datasets

Figure 4: Parameters estimates of under-dispersed datasets



a) High mean equi-dispersed datasets

a) Moderate mean equi-dispersed datasets

a) Low mean equi-dispersed datasets

Figure 5: Parameters estimates of equi-dispersed datasets

4.3 PREDICTION BIAS

The bias of an estimator is defined as the difference between an estimator's expected value and the true value of the parameter being estimated. If the bias is zero then the estimator is said to be unbiased. The bias of an estimator $\hat{\theta}$ is calculated as $E(\hat{\theta}) - \theta$ where θ is the true value of the parameter and the estimator $\hat{\theta}$ is a function of the observed data.

The bias of β and α parameters is calculated as the difference between their average values from 5 samples and the true (or assigned) value in each scenario.

The bias of centering parameter *'μ'* is calculated as

$$E(\hat{\mu}) - \mu = (\overline{\hat{\beta}_0} - \beta_0) + (\overline{\hat{\beta}_1} - \beta_1)\overline{X}_1 + (\overline{\hat{\beta}_2} - \beta_2)\overline{X}_2 \qquad (15)$$

The bias of centering parameter *'ν'* is calculated as

$$E(\hat{\nu}) - \nu = (\overline{\hat{\alpha}_0} - \alpha_0) + (\overline{\hat{\alpha}_1} - \alpha_1)\overline{X}_1 + (\overline{\hat{\alpha}_2} - \alpha_2)\overline{X}_2 \qquad (16)$$

Where $(\overline{\hat{\beta}_i} - \beta_i)$ and $(\overline{\hat{\alpha}_i} - \alpha_i)$ are the bias in the parameters and $\overline{X}_i$ is the average value of the independent variable, which is typically 0.5 since the independent variables are randomly simulated between 0 and 1. As seen from Figure 6, with the exception of the under-dispersed data, the bias increased as the mean values decreased. The bias did not change significantly for the under-dispersed data from one mean to other. The bias becomes worse for the over-dispersed datasets at low mean values.

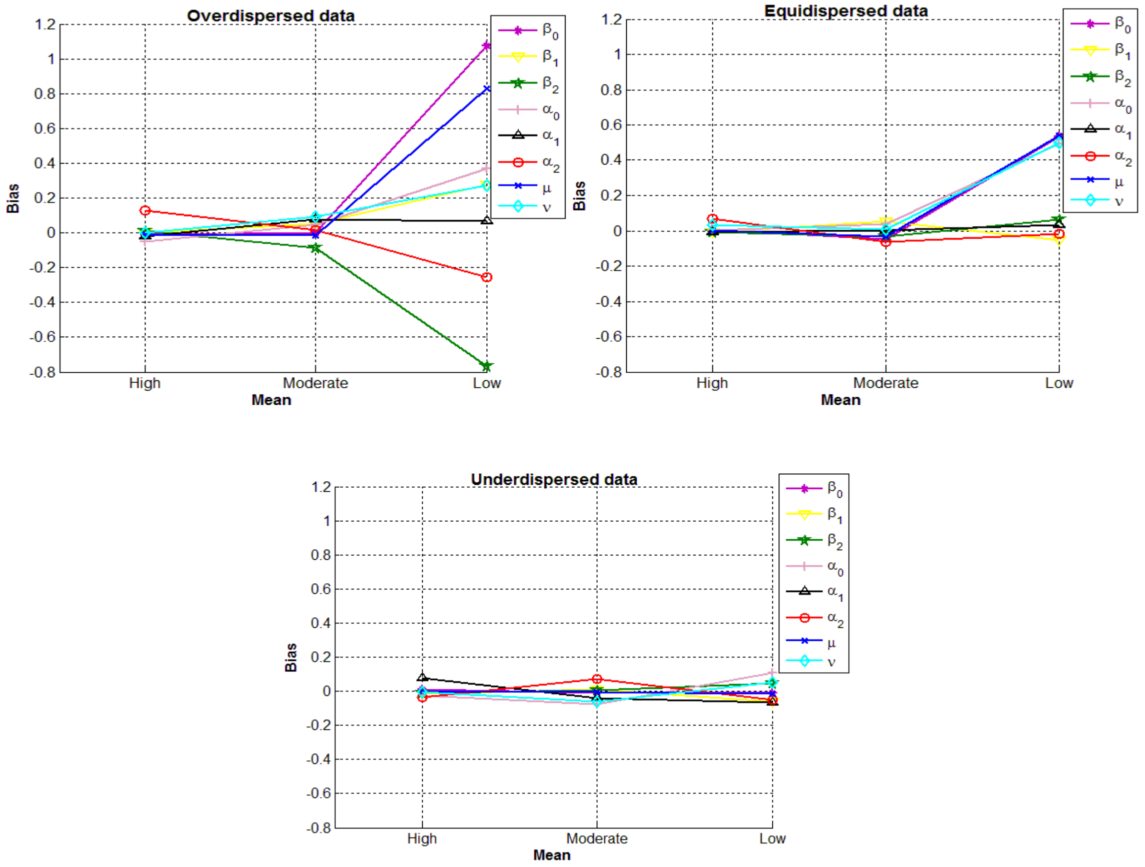
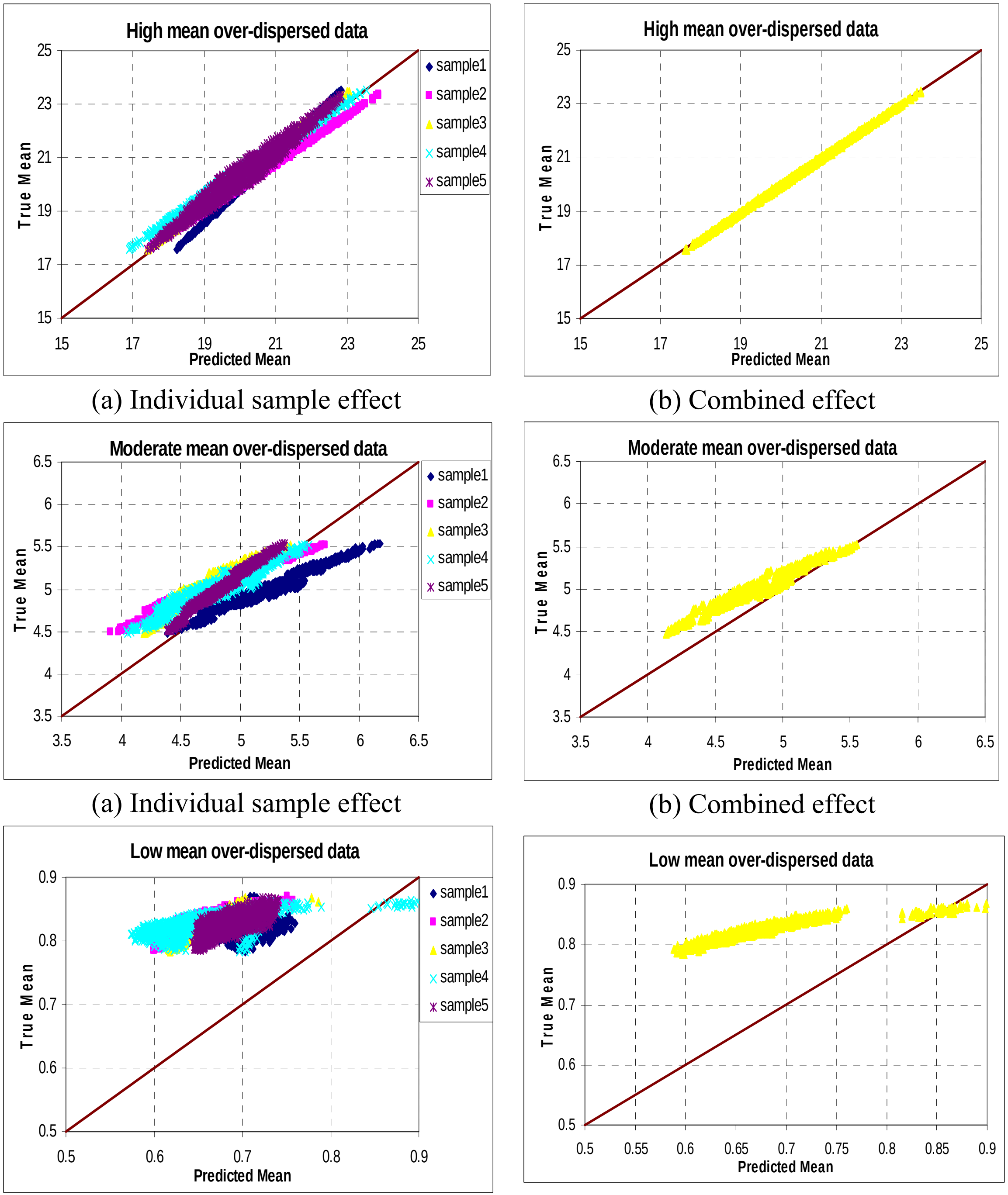


Figure 6: Prediction bias of the parameters

The biases in the mean values are plotted and are shown in Figures 7-9. Figure 7 gives plots of the estimated and true mean values for the over-dispersed datasets. For the true mean, first the true μ and ν parameters were calculated from the true (or assigned) parameters. We then simulated 100,000 random counts from the COM-Poisson distribution for the given μ and ν. The mean of these random variables gives the true mean for each sample. Similarly, the predicted μ and ν parameters were calculated from the estimated parameters for each of the five samples. Again, 100,000 random counts were simulated from the COM-Poisson distribution for the given μ and ν. The mean of these random variables gives the predicted mean for each sample. The second subplot corresponds to the combined effect of all five samples. Instead of the parameters estimated for each sample, the average of the estimated parameter is considered in calculating the predicted mean in these plots.

The COM-Poisson distribution performs better for high and moderate mean for all three categories of dispersion. For over-dispersed and equi-dispersed datasets, the performance is worse for all low sample mean values. The COM-Poisson distribution works well for all sample mean values for the under-dispersed datasets. However, another study conducted by the authors (Geedipally et al., 2008) related to the application of COM GLM for analyzing motor vehicle crash data exhibiting under-dispersion (conditional on the mean) showed that the estimated mean is an unreliable estimate of traffic crashes at extremely low sample mean values (~0.3) for v >1. The centering parameter of the distribution was itself found to be a preferable estimate for predicting crashes.



(a) Individual sample effect

(b) Combined effect

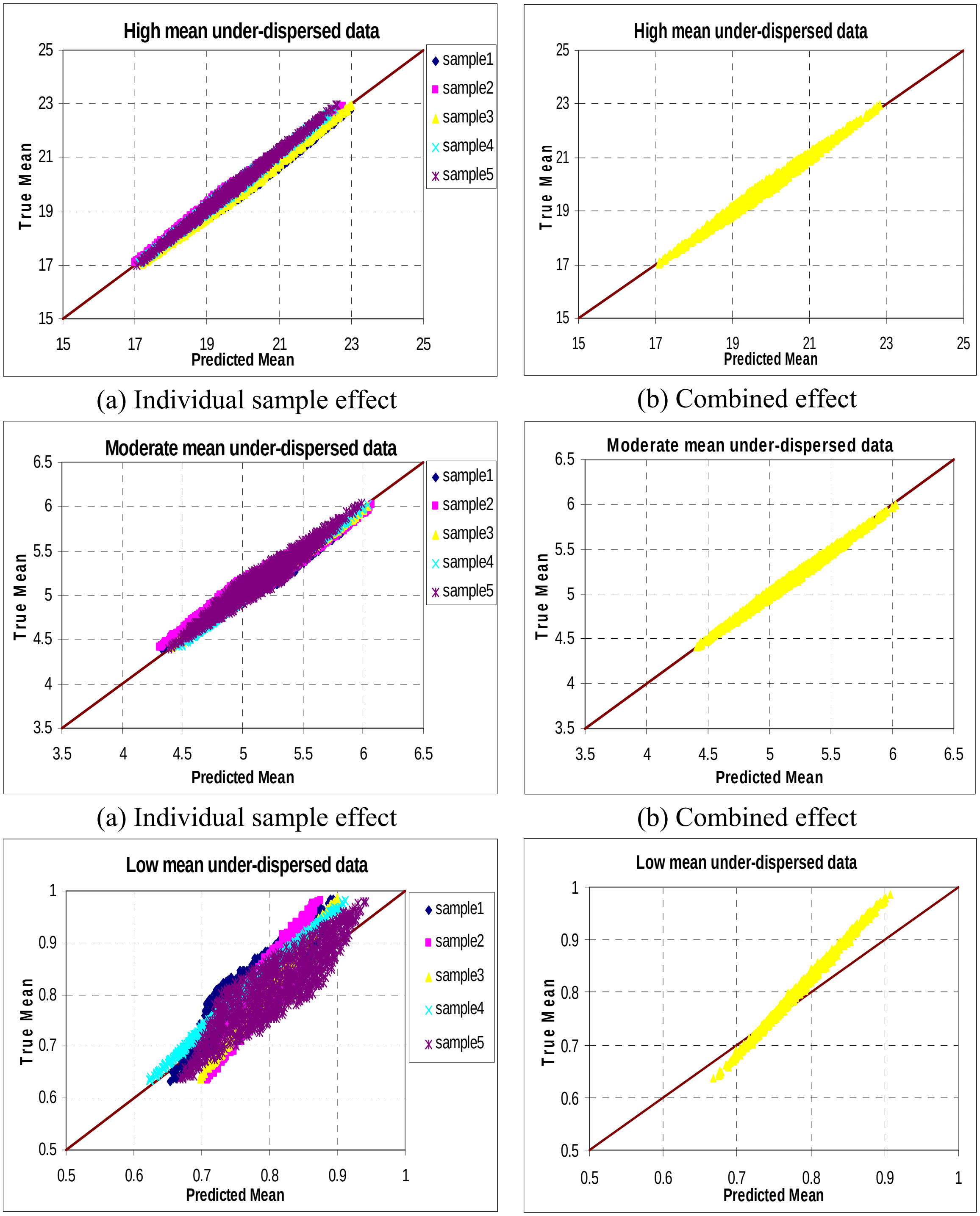
(a) Individual sample effect

(b) Combined effect

(a) Individual sample effect

(b) Combined effect

Figure 7: Prediction accuracy for over-dispersed datasets



(a) Individual sample effect

(b) Combined effect

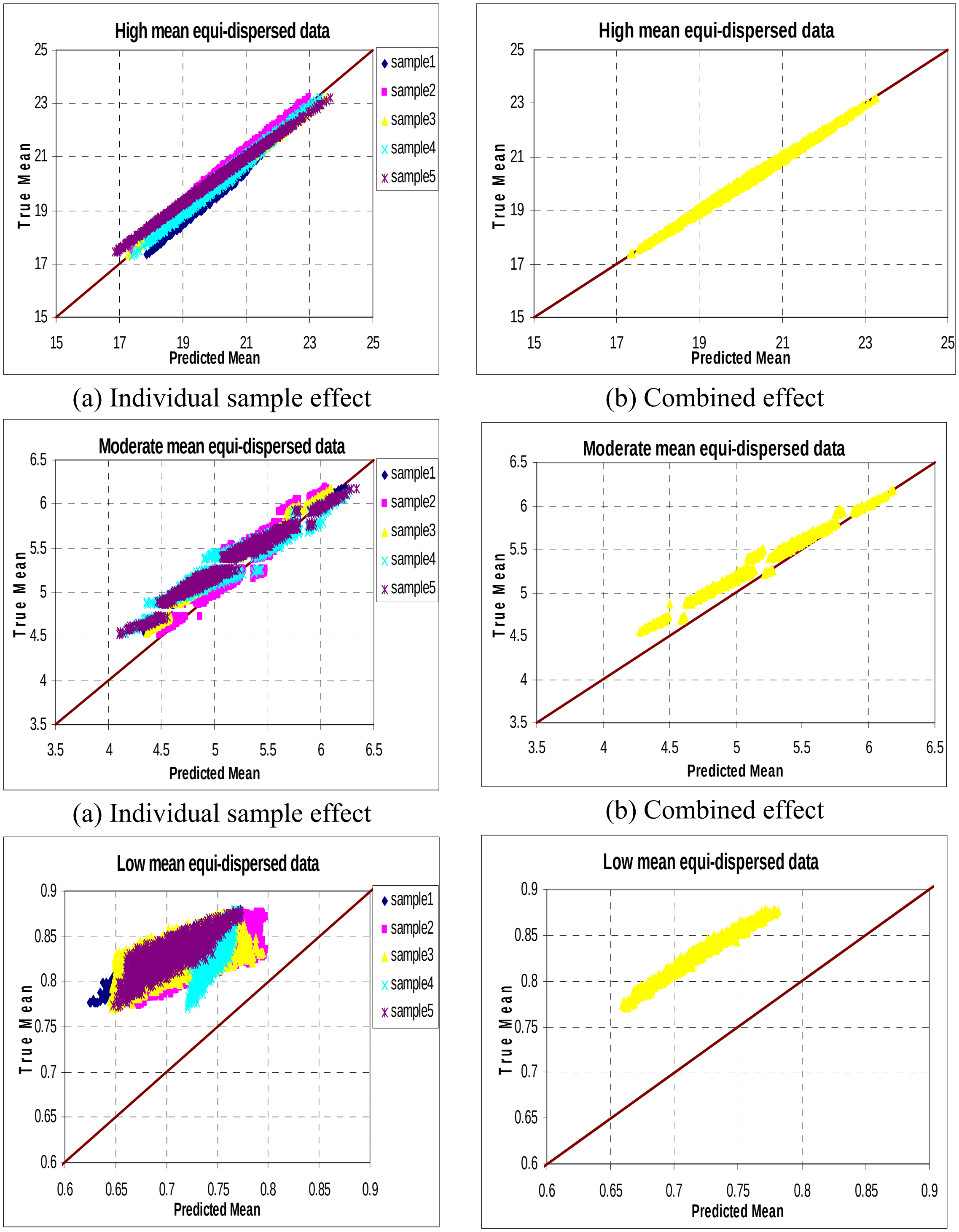
(a) Individual sample effect

(b) Combined effect

(a) Individual sample effect

(b) Combined effect

Figure 8: Prediction accuracy for under-dispersed datasets



(a) Individual sample effect

(b) Combined effect

(a) Individual sample effect

(b) Combined effect

(a) Individual sample effect

(b) Combined effect

Figure 9: Prediction accuracy for equi-dispersed datasets

4.4 ACCURACY OF THE ASYMPTOTIC MEAN APPROXIMATION

The centering parameter μ is believed to adequately approximate the mean when μ >10 based on the asymptotic approximation developed by Shmueli et al. (2005). However, the deviation of μ for mean values below 10 (μ < 10) has not been investigated. We chose one sample from each of the nine scenarios as a basis for estimating the accuracy of the asymptotic mean approximation. First, the μ and v parameters were calculated from the estimated parameters. We examined the goodness of this approximation by simulating 100,000 random values from the COM-Poisson for a given μ and v. We then plotted the mean of the simulated values against the asymptotic mean approximation ($E\left[Y\right] \approx \mu + 1/2\nu - 1/2$). The results showed that the asymptotic mean approximates the true mean accurately even for 10 > $E[Y]$ > 5. As the sample mean value decreases below 5, the accuracy of the approximation drops. As seen in Figure 10, the asymptotic approximation holds well for all datasets with high and moderate mean values irrespective of the dispersion in the data. The approximation is also accurate for low sample mean values for under-dispersed datasets. The accuracy drops significantly for the low sample mean values for over-dispersed and equi-dispersed datasets. There is not much difference between the asymptotic mean approximation and the true mean for $E[Y]$ > 10. It starts to deviate at the moderate mean values although the difference is not high. The difference can clearly be observed for the low sample mean values, particularly for the over-dispersed and equi-dispersed datasets. This shows that one must be careful in using the asymptotic approximation for the mean of the COM GLM to estimate future event counts for datasets characterized by low sample mean values.

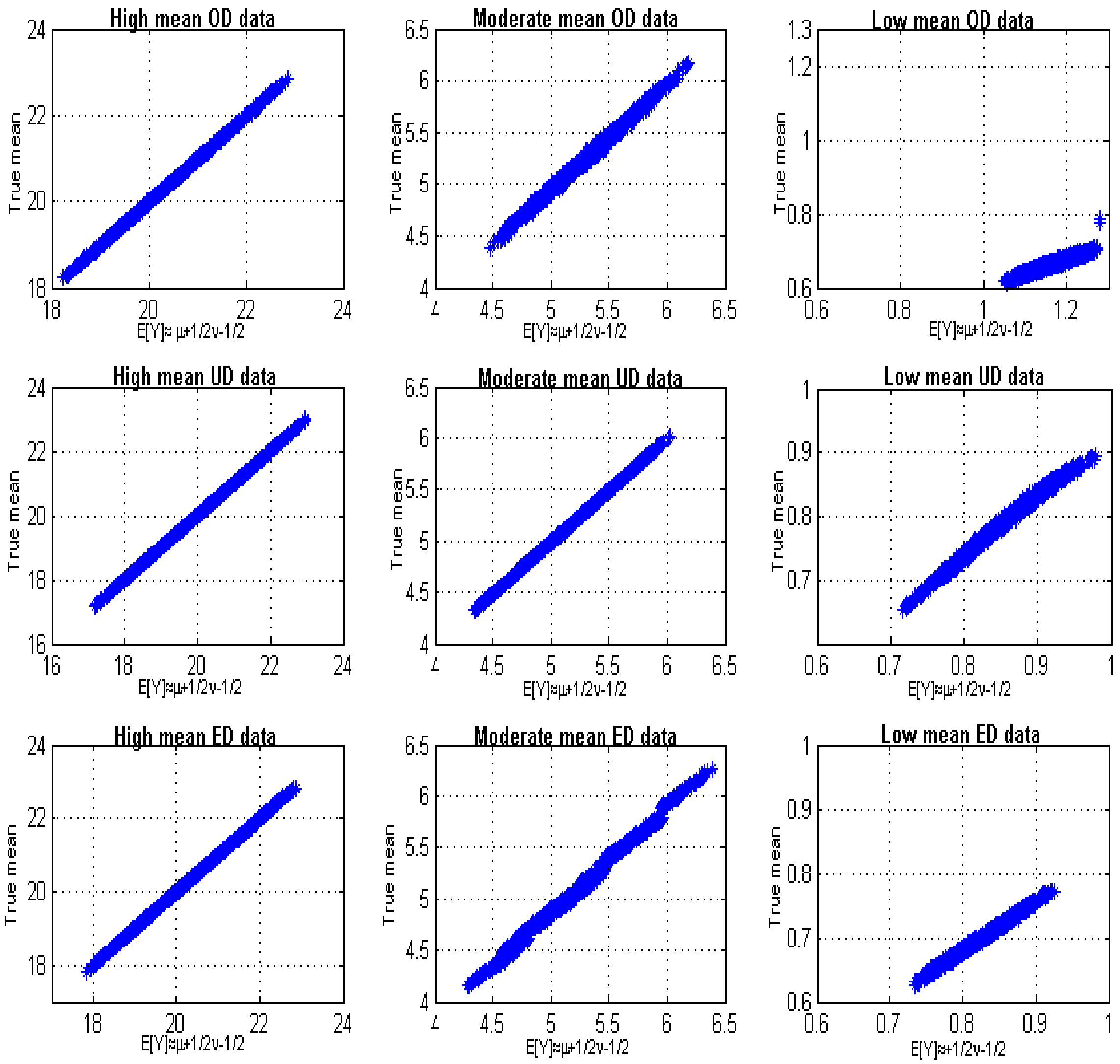


Figure 10: Mean versus asymptotic approximation

4.5 COMPUTATIONAL TIME

We also investigated the computational time needed for the WinBUGS MCMC implementation of the COM GLM. We ran the COM-Poisson MCMC model on a computer with a 1.5GHz Pentium 4 processor and 512 MB of RAM. Each run consisted of 3 chains of 50,000 replications each. The computational times for all of the datasets were plotted against the mean values of the counts in those datasets and are shown in the Figure 11. Datasets with higher sample means required more computational time for a given number of replications than datasets with low sample mean. This is mainly attributable to the convergence of the *S* term. The centering parameter causes the numerator to be large for high sample mean values, requiring that more terms be included in the approximation to achieve suitable convergence of the approximation of the series. Also, it is important to note that the over-dispersed datasets required more computational time than the other type of datasets. The shape parameter plays vital role in the convergence of the *S* term. Lower values for the shape parameter require higher amounts of computational time.

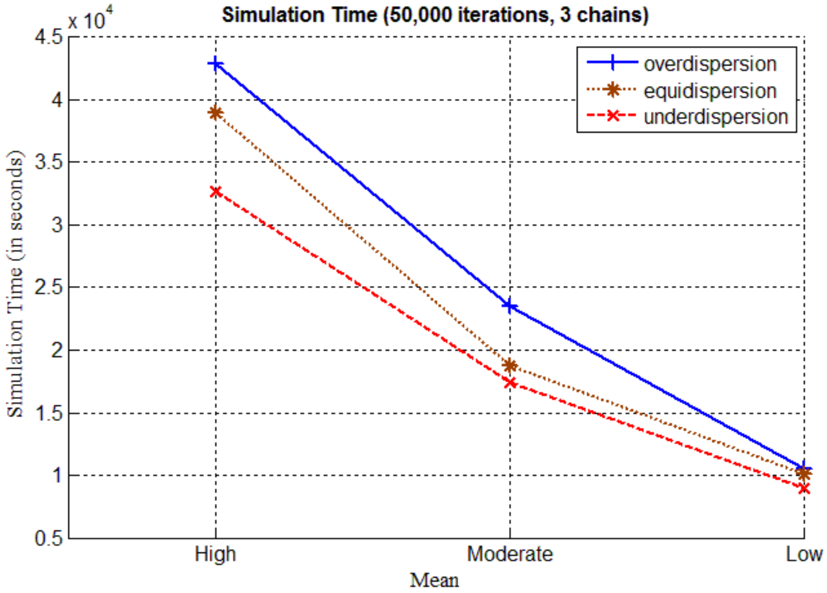


Figure 11: Computational time for the WinBUGS MCMC implementation

5. DISCUSSION

This paper shows that the COM GLM is flexible in handling count data irrespective of the dispersion in the data. First, the true parameters lie in the 95% credible interval for nearly all cases and are generally close to the estimated posterior mean of the parameters. The credible intervals were found to be wider for the low mean values for both the centering and shape parameters. The bias in the prediction of the parameters and the mean also increases as the data sample mean values decreases. Even at the low sample mean values, the bias is considerably less for under-dispersed datasets than for over-dispersed and equi-dispersed datasets. Despite its flexibility in handing count data with all dispersions, the COM-Poisson distribution suffers from important limitations for low mean over-dispersed data. This similar behavior is also exhibited by Negative Binomial (Poisson-gamma) models (Lord, 2006). Second, the asymptotic approximation of the mean suggested by Shmueli et al (2005) approximates the true mean adequately for $E[Y] > 5$. This value found through numerical analysis of the COM GLM is substantially lower than the lower bound value of 10 suggested by Shmueli et al. (2005). As the sample mean value decreases, the accuracy of the approximation becomes lower. The asymptotic approximation is accurate for all datasets with high and moderate sample mean values irrespective of the dispersion in the data. The approximation is also accurate for low sample mean values for under-dispersed datasets. However, the accuracy drops substantially for low sample mean values for over-dispersed and equi-dispersed datasets. Third, datasets with higher sample mean values required more computational time for a given number of replications than the low mean datasets did. Similarly, it is important to note that the over-dispersed datasets required more computational time than the other type of datasets.

6. CONCLUSIONS

This paper has documented the performance of COM GLM for datasets characterized by different variances and sample mean values. The results of this study showed that the COM GLMs can handle under-, equi- and over-dispersed datasets with different mean values, although the credible intervals are found to be wider for low sample mean values. Despite its limitations for low sample mean values for over-dispersed datasets, the COM GLM is still a flexible method for analyzing count data. The asymptotic approximation of the mean is accurate for all datasets with high and moderate sample mean values irrespective of the dispersion in the data, and it is also accurate for low sample mean values for under-dispersed datasets. Furthermore, the computational effort required for the MCMC implementation of the COM GLM is not prohibitive. Finally, the COM GLM is a promising, flexible regression model for count data.

ACKNOWLEDGEMENTS

This work was partially supported by NSF grant number ECCS-0725823 to Seth Guikema.

REFERENCES

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, 564–572.

Borle, S., Boatwright, P. and Kadane.J.B., 2006. The Timing of Bid Placement and Extent of Multiple Bidding: An Empirical Investigation Using eBay Online Auctions. Statistical Science Vol 21(No. 2): 194-205.

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

Geedipally, S., Lord, D., and Guikema, S., Extension of the application of Conway-Maxwell-Poisson Models: Analyzing Traffic Crashes Exhibiting Underdispersion. Working paper (Texas A&M University, College Station, TX, 2008).

Guikema, S.D. and Coffelt J.P., 2008. A flexible count data regression model for risk analysis. Risk Analysis, Vol. 28, No. 1, pp. 213-223.

Kadane, J.B., Shmueli, G., Minka, T.P., Borle, S. and Boatwright, P., 2006. Conjugate analysis of the Conway-Maxwell-Poisson distribution. Bayesian Analysis, Vol. 1, pp. 363-374.

Lord, D., 2006. Modeling Motor Vehicle Crashes using Poisson-gamma Models: Examining the Effects of Low Sample Mean Values and Small Sample Size on the Estimation of the Fixed Dispersion Parameter. Accident Analysis & Prevention, Vol. 38, No. 4, pp. 751-766.

Lord, D., Guikema, S.D. and Geedipally, S., 2008. Application of the Conway-Maxwell-Poisson Generalized Linear Model for Analyzing Motor Vehicle Crashes. Accident Analysis & Prevention, in press.

Luc Devroye, 1986. Non-uniform Random Variate Generation. Springer-Verlag, New York

Ridout, M.S. and Besbeas, P., 2004. An empirical model for underdispersed count data. Statistical Modelling, 4: 77–89.

Sellers, K.F. and Shmueli, G., 2008. A Flexible Regression Model for Count Data. Working paper, Georgetown Univerity, Washington, DC.

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

Spiegelhalter, D.J., Thomas, A., Best, N.G., Lun, D., 2003. WinBUGS Version 1.4.1 User Manual. MRC Biostatistics Unit, Cambridge. Available from: <http://www.mrcbsu.cam.ac.uk/bugs/welcome.shtml>.

Telang, R., Boatwright, P. and Mukhopadhyay,T., 2004. A Mixture Model for Internet Search-Engine Visits. Journal of Marketing. 41 (May), 206-214.