Studying the Third Cumulant of the Mixture of Dirichlet-multinomial Distributions

Farzana Afroz
Department of Statistics, University of Dhaka, Dhaka-1000, Bangladesh
(Received : 12 July 2021 : Accepted : 14 October 2021)

Abstract
Traditionally, the overdispersion parameter \( \phi \) is estimated by using Pearson’s lack of fit statistic \( X^2 \) or the Deviance statistic \( D \), which do not perform well in the case of sparse data. This paper particularly focuses on an estimator \( \hat{\phi}_{\text{new}} \) of overdispersion parameter which was proposed for sparse multinomial data. The estimator was derived on the basis of an assumption on the 3rd cumulant of the response variable. When the data comes from the Dirichlet-multinomial distribution \( \hat{\phi}_{\text{new}} \) is known to have the lowest root mean squared error comparing to the other three estimators. In this paper the 1st to 3rd order raw moments of the finite mixture of Dirichlet-multinomial distributions are derived, which results in complicated mathematical expressions. Furthermore, it is found that the 3rd cumulant of this mixture does not satisfy the assumption which is considered in the derivation of \( \hat{\phi}_{\text{new}} \).

Keywords: multinomial distribution, overdispersion, Dirichlet-multinomial, finite-mixture, moments, cumulant.

I. Introduction
The correlation among the observations making up the counts of multinomial data creates overdispersion. Ignoring overdispersion may lead to the serious under-estimation of the standard errors and consequently the model parameters will be incorrectly tested. Overdispersed multinomial data may arise in many areas, such as mark-recovery and mark-recapture modelling, household surveys, DNA sequence analysis, hyperspectral image (HSI) classification. Several approaches can be found in the literature for handling multinomial data that exhibit overdispersion. The most common likelihood approaches for modelling multinomial overdispersed data are Dirichlet-multinomial distribution (Mosimann\(^1\)) and Finite mixture distribution (Morel and Nagaraj\(^2\)). Another approach is the quasi-likelihood (QL) method introduced by Wedderburn\(^3\) and McCullagh\(^4\) and Nelder\(^5\), and generalized estimating Equations (GEE) by Liang and Zeger\(^6\) and Zeger and Liang\(^7\). The quasi-likelihood method is simple to apply as only the first two moments of the response variables need to be specified. Also, the maximum likelihood estimate is the same as the maximum quasi likelihood estimate for GLMs. In quasi-likelihood method it is assumed that \( \text{Var}(Y) = \phi V \), where \( V \) is the variance function, \( \phi > 1 \) indicates overdispersion. The classical estimators of overdispersion are

\[
\hat{\phi}_p = \frac{X^2}{n - p} \text{ or } \hat{\phi}_D = \frac{D}{n - p},
\]

where \( X^2 \) is the Pearson’s lack of fit statistics and \( D \) is the residual deviance, \( n \) is the total number of observations and \( p \) is the number of parameters estimated. Asymptotically both \( X^2 \) and \( D \) follow \( \chi^2 \) distribution with \( n - p \) degrees of freedom. However, the asymptotic results of these test statistics depends on the total counts being sufficiently large, which is unlikely to be the case for sparse data containing many small counts.

In 1986 McCullagh\(^7\) argued that for assessing goodness of fit the conditional distribution of the test statistic is more relevant than the marginal distribution. For sparse discrete data McCullagh\(^7\) derived conditional moments of \( X^2 \) and \( D \).

Farrington\(^8\) used an estimating equations approach to extend the results of McCullagh\(^7\) to models with any type of link function.

Fletcher\(^9\) considered the problem of estimating the overdispersion parameter \( \phi \) when fitting a generalized linear model to sparse data. He proposed a new estimator of \( \phi \) that has a smaller variance than Wedderburn’s and Farrington’s estimator, subject to a condition on the third cumulant (\( \kappa_3 \)) of the response variable. Under the assumption \( \kappa_3 = \alpha \kappa_3^0 \), where \( \alpha > \phi^2 \) and \( \kappa_3^0 \) is the third cumulant of the Poisson distribution, Fletcher\(^9\) showed that that

\[
\text{var}(\hat{\phi}_{\text{new}}) \leq \text{var}(\hat{\phi}_F) < \text{var}(\hat{\phi}_P)
\]

where, \( \hat{\phi}_F \) is the Farrington’s estimator. Therefore, the assumption on the 3rd cumulant is needed to justify that the new estimator is more efficient compared to the other estimators. Through simulation study Fletcher\(^9\) showed that the proposed estimator has the lowest level of root mean squared error (RMSE) for the increasing level of overdispersion compared to the Wedderburn’s and Farrington’s estimator when the data comes from Negative binomial distribution and Neyman Type A distribution.

Considering the same assumption on the 3rd cumulant as Fletcher\(^9\), Afroz\(^10\) proposed an estimator \( \hat{\phi}_{\text{new}} \) for multinomial data, which can be defined as follows:

suppose there are \( n \) independent multinomial random variables \( Y_i = (Y_{i1}, Y_{i2}, ..., Y_{ik}) \) having mean vector \( \mu_i = (\mu_{i1}, \mu_{i2}, ..., \mu_{ik})' \), then

\[
\hat{\phi}_{\text{new}} = \frac{\hat{\phi}_p}{1 + \tilde{\xi}},
\]

where \( \tilde{\xi} = \frac{\sum_{i,j}(y_{ij} - \bar{y}_{ij})^2}{n(k-1)} \) and \( \bar{y}_{ij} \) is the quasi likelihood estimate of \( \mu_{ij} \). Also, Farrington’s estimator for multinomial data can be defined as

\[
\hat{\phi}_F = \hat{\phi}_p - \frac{n(k-1)^{-1}}{n(k-1) - \tilde{\xi}}
\]

\(^\text{Author for correspondence. e-mail: fafroz87@du.ac.bd}\)
Afroz\textsuperscript{10} showed that $\hat{\varphi}\text{new}$ has the lowest RMSE comparing to $\hat{\varphi}_p$, $\hat{\varphi}_d$ and $\hat{\varphi}_F$, when the data are highly sparse and follow Dirichlet-Multinomial distribution. However, it is shown that $\hat{\varphi}\text{new}$ did not performed the best when the data are generated from the finite mixture of Dirichlet-multinomial distributions. Afroz\textsuperscript{10} derived the 3rd cumulant of the Dirichlet-multinomial distribution and showed that it satisfies the assumption discussed in Fletcher\textsuperscript{9}. The motivation of this paper comes from a question raised in Afroz\textsuperscript{10}, whether the 3rd cumulant of the mixture of Dirichlet-multinomial distributions satisfy the assumption or not. In this paper, the 1st to 3rd order raw moments of the mixture of Dirichlet-multinomial distribution sare derived, also using the raw moments the 3rd cumulant is derived.

II. Dirichlet-multinomial Distribution

Let $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ik-1})'$ denote the observations from a typical cluster of size $m$. Here $Y_{ij}$ denotes the count in cluster $i$ and category $j$ ($i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, k - 1$) and $Y_k = m - (Y_{i1} + Y_{i2} + \cdots + Y_{ik-1})$. The Dirichlet-multinomial distribution proposed by Mosimann\textsuperscript{1} provides a way to model categorical data exhibiting overdispersion. Here the overdispersion is believed to arise from the fact that the cell probabilities $\pi = (\pi_1, \pi_2, \ldots, \pi_{k-1})'$ vary randomly according to a Dirichlet distribution, with probability density function

$$f(\pi_i) = \frac{\Gamma(c)}{\prod_{j=1}^{k-1} \Gamma(cp_{ij})} \prod_{j=1}^{k} \pi_i^{cp_{ij}-1}$$

where $c = \frac{1+r}{2}, 0 < r < 1$, and $\Gamma(\cdot)$ denotes the gamma function. Thus it is assumed that $Y_i|\pi_i$ is distributed as a k-dimensional multinomial random variable and the marginal distribution of $Y_i$ is then the Dirichlet-multinomial, with probability mass function

$$P(Y_i = y_i) = \frac{m!}{y_{i1}! y_{i2}! \cdots y_{ik}!} \frac{\Gamma(c)}{\prod_{j=1}^{k} \Gamma(cp_{ij})} \prod_{j=1}^{k} \pi_i^{c}.$$

The mean and variance of $Y_i$ are as follows

$$E(Y_i) = m\pi_i$$

and

$$\text{var}(Y_i) = \{1 + \tau(m - 1)\} m\{\text{diag}(\pi_i) - \pi_i\pi'_i\}$$

where $\text{diag}(\pi_i)$ is a diagonal matrix with diagonal elements $\pi_{i1}, \pi_{i2}, \ldots, \pi_{i(k-1)}$. The term $\{1 + \tau(m - 1)\}$ indicates extra variation comparing to the usual covariance of a multinomial distribution. Note that for $\tau = 0$ the Dirichlet-multinomial distribution and the usual multinomial distribution have common covariance matrix. Newcomer et al.\textsuperscript{11} derived the higher order moments of the two commonly used multinomial overdispersion models. Following the results of Newcomer et al.\textsuperscript{11} the first through third order moments of Dirichlet-multinomial distribution are

I. $E(Y_{ij}) = mp_{ij}$,

II. $E(Y_{ij}^2) = m(m - 1)\frac{(cp_{ij}+1)}{(c+1)} + mp_{ij}$,

III. $E(Y_{ij}^3) = m(m - 1)(m - 2)\frac{(cp_{ij}+1)(cp_{ij}+2)}{(c+1)(c+2)}p_i + 3m(m - 1)\frac{(cp_{ij}+1)}{(c+1)} + mp_{ij}$.

III. Example

In order to examine the numerical differences between the estimators, mark-recovery data on herring gulls (Larus argentatus) from Kent Island in Canada are used here. From 1934 to 1939, 31,694 fledging gulls were banded, of which 1099 were recovered after death. The detailed description can be found in Paynter\textsuperscript{12}. Suppose, $s$ is the probability that a bird survives a year and $r$ is the probability that a banded dead bird is reported. Simple form of Seber\textsuperscript{3} model is used to model cell probabilities of the data. After fitting product multinomial model $r$ and $s$ are estimated as 0.035 and 0.655 respectively. In order to, construct a suitable confidence interval of the parameters, $r$ and $s$ a proper estimator of $\varphi$ is needed. Now a simulation study is performed to select the best estimator of $\varphi$ by generating data from Dirichlet-multinomial distribution. The recovery and survival probabilities were set to the estimated value and $\varphi$ is varied from 2 to 5 in the simulation study. Now the root mean squared error for all the 4 estimators are calculated for each level of $\varphi$. From the results displayed in Table 1 it is apparent that, $\hat{\varphi}\text{new}$ outperforms the other estimators for different levels of $\varphi$. It is noticeable that RMSE increases for the increasing values of $\varphi$, however $\hat{\varphi}\text{new}$ has the lowest RMSE among all the estimators in each case.

<table>
<thead>
<tr>
<th>$\varphi$</th>
<th>$\hat{\varphi}_p$</th>
<th>$\hat{\varphi}_d$</th>
<th>$\hat{\varphi}_F$</th>
<th>$\hat{\varphi}\text{new}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3.112</td>
<td>2.167</td>
<td>1.203</td>
<td>0.634</td>
</tr>
<tr>
<td>3</td>
<td>9.879</td>
<td>7.418</td>
<td>1.967</td>
<td>1.144</td>
</tr>
<tr>
<td>4</td>
<td>5.085</td>
<td>3.527</td>
<td>2.759</td>
<td>1.584</td>
</tr>
<tr>
<td>5</td>
<td>11.576</td>
<td>9.434</td>
<td>3.581</td>
<td>2.175</td>
</tr>
</tbody>
</table>

Therefore, from the simulation results it is clear that $\hat{\varphi}\text{new}$ should be used to construct the confidence interval of the parameters.

IV. Mixture of Dirichlet-multinomial Distributions

Suppose that the multinomial data $Y = (Y_{11}, \ldots, Y_{1k})$ is collected from a population which is a mixture of two groups having different parameters $\pi_1 = (\pi_{11}, \pi_{12}, \ldots, \pi_{1k})$ and $\pi_2 = (\pi_{21}, \pi_{22}, \ldots, \pi_{2k})$. Each observation comes from group-1 and group-2 with probability $w_1$ and $w_2$ respectively, where $w_1 + w_2 = 1$. Let us consider the following setup with subscript $i$ removed for simplicity

$$Y = Y_1 + Y_2$$

$$Y_1 = (Y_{11}, \ldots, Y_{1k}), \quad Y_2 = (Y_{21}, \ldots, Y_{2k})$$

where

$$Y_i|m_i, \pi_i \sim \text{Multinom}(m_i, \pi_i).$$
\[ Y_2|m_2, \pi_2 \sim \text{Multinom}(m_2, \pi_2), \]

\[ m_1 + m_2 = m, \]

and

\[ \pi_1 \sim \text{Dirichlet}(c_1 p_{11}, \ldots, c_1 p_{1k}), \]

\[ \pi_2 \sim \text{Dirichlet}(c_2 p_{21}, \ldots, c_2 p_{2k}). \]

We then have

\[ E(\pi_{1j}) = \frac{c_1 p_{1j}}{\sum_j c_1 p_{1j}}, \quad E(\pi_{2j}) = \frac{c_2 p_{2j}}{\sum_j c_2 p_{2j}}, \]

\[ \text{Var}(\pi_{1j}) = \pi_{1j}(1 - \pi_{1j}), \quad \text{Var}(\pi_{2j}) = \pi_{2j}(1 - \pi_{2j}). \]

Now following the results of Afroz\(^1\) the expected value and variance of \( Y_j \) can be obtained as follows

\[ \mu'_1 = E_{m, \pi_1, \pi_2}(Y_1 + Y_2) \]

\[ = E_m \left( E_{\pi_1,Y_2|m}(Y_1 + Y_2) \right) \]

\[ = E_m \left( E_{\pi_1}(E_{Y_2|m}(Y_1 + Y_2)) \right) \]

\[ = E_m \left( E_{\pi_1}(m_1 \pi_{1j} + m_2 \pi_{2j}) \right) \]

\[ = m (w_1 p_{1j} + w_2 p_{2j}) \]

Similarly

\[ \mu_2 = \text{Var}_{m, \pi_1, \pi_2}(Y_1 + Y_2) \]

\[ = E_m \left( \text{Var}_{\pi_1,Y_2|m}(Y_1 + Y_2) \right) \]

\[ + \text{Var}_m \left( E_{\pi_1,Y_2|m}(Y_1 + Y_2) \right) \]

\[ = E_m \left( m_1(1 + ((m_1 - 1) \tau_1) p_{1j}(1 - p_{1j}) 
+ m_2(1 + ((m_2 - 1) \tau_2) p_{2j}(1 - p_{2j}) \right) \]

\[ + \text{Var}_m \left( m_2 p_{2j} + m_2 (m_2 - 1) \pi_{2j}(1 - \pi_{2j}) \right) \]

\[ = (1 + (m - 1) \tau_1 w_1) mw_1 p_{1j}(1 - p_{1j}) \]

\[ + (1 + (m - 1) \tau_2 w_2) mw_2 p_{2j}(1 - p_{2j}) \]

\[ + p_{1j}^2 w_1(1 - w_1) + p_{2j}^2 w_2(1 - w_2) \]

\[ - 2mw_1 w_2 p_{1j} p_{2j} \]

(1)

Now the second order raw moment \( \mu'_2 \) of \( Y_j \) can be calculated by using Equations (1) and (2). For the simplified form of \( \mu'_2 \), the software Mathematica\(^14\) Version 12.1 is used. The final form of \( \mu'_2 \) is as follows

\[ \mu'_2 = \mu_2 + \mu_2^2 = -mp_{1j} w_1 (p_{1j} w_1 (1 + (m - 1) \tau_1) 
+ \tau_1 w_1 (1 - m - 1) - 2mp_{1j} p_{2j} w_2 (m_1 \pi_1 w_1 + m_2 \pi_2 w_2) 
-m_2 p_{2j} w_2 (m_2 - 1) \tau_2 + 1) 
+ (1 - m) \tau_2 w_2 (1) + (m - 1) p_{1j} \tau_1 - p_{1j}^2 (w_1 - 1 - \tau_1 w_1 + m \tau_1 w_1) \]

\[ - 2mp_{1j} p_{2j} w_2 \]

\[ + (mp_{1j} w_1 + \]

\[ mp_{2j} w_2)^2 + mw_2 \left( (1 + m - 1) \tau_2 p_{2j} w_2 - p_{2j}^2 (w_2 (1 - (1 - m) \tau_2) - 1) \right) \]

(3)

In order to derive 3rd order raw moment \( \mu'_3 \) of \( Y_j \), results from Newcomer et al\(^11\) are used. The derivation of \( \mu'_3 \) is as follows

\[ \mu'_3 = E_{m, \pi_1, \pi_2}(Y_1 + Y_2)^3 \]

\[ = E_{m, \pi_1, \pi_2}(Y_1^3 + 3Y_1 Y_2^2 + 3Y_1^2 Y_2 + 3Y_2^3) \]

\[ = E_m \left( E_{\pi_1}(E_{Y_1|m}(Y_1^3 + 3Y_1 Y_2^2 + 3Y_1^2 Y_2 + 3Y_2^3)) \right) \]

\[ = E_m \left( m_1 \pi_{1j}(1 - 3\pi_{1j} + 3m_1 \pi_{1j} + 2\pi_{1j}^2 - 3m_1 \pi_{1j}^2) \right) \]

\[ + 3 \left( m_2 p_{2j}(1 - \pi_{2j} + m_2 \pi_{2j}) \right) m_2 \pi_{2j} \]

\[ + 3 \left( m_2 \pi_{2j} + m_2 \pi_{2j}^2 \right) \right) \]

\[ = E_m \left( m_1 p_{1j} + 3m_1 (m_1 - 1) \frac{(C_1 p_{1j} + 1) p_{1j}}{(C_1 + 1)} \right) \]

\[ + m_1 (m_1 - 1) (m_2 - 1) \frac{(C_1 p_{1j} + 2) p_{1j}}{(C_1 + 1)(C_1 + 2)} \]

\[ + 3 \left( m_1 p_{1j} + 3m_1 (m_1 - 1) \frac{(C_1 p_{1j} + 1) p_{1j}}{(C_1 + 1)} \right) \]

\[ + 3 \left( m_2 p_{2j} + m_2 (m_2 - 1) \frac{(C_2 p_{2j} + 1) p_{2j}}{(C_2 + 1)} \right) \]

\[ + m_2 (m_2 - 1) (m_2 - 2) \frac{(C_2 p_{2j} + 2) p_{2j}}{(C_2 + 1)(C_2 + 2)} \]

\[ = mw_1 p_{1j} + 3m_1 (m_1 - 1) w_1^2 \frac{(C_1 p_{1j} + 1) p_{1j}}{(C_1 + 1)} \]

\[ + m_1 (m_1 - 1) (m_2 - 1) w_2^2 \frac{(C_1 p_{1j} + 2) p_{1j}}{(C_1 + 1)(C_1 + 2)} \]

\[ + 3 \left( m_1 p_{1j} + m_1 (m_1 - 1) \frac{(C_1 p_{1j} + 1) p_{1j}}{(C_1 + 1)} \right) \]

\[ + 3 \left( mw_2 p_{2j} + m_1 (m_1 - 1) w_2^2 \frac{(C_1 p_{1j} + 1) p_{1j}}{(C_1 + 1)} \right) \]

\[ + m_1 (m_1 - 1) (m_2 - 1) w_2^2 \frac{(C_1 p_{1j} + 2) p_{1j}}{(C_1 + 1)(C_1 + 2)} \]

\[ + m_1 (m_1 - 1) (m_2 - 2) w_2^2 \frac{(C_1 p_{2j} + 1) p_{2j}}{(C_1 + 1)(C_1 + 2)} \]

\[ + m_2 (m_2 - 1) (m_2 - 2) w_2^2 \frac{(C_2 p_{2j} + 2) p_{2j}}{(C_2 + 1)(C_2 + 2)} \]

(4)

Note that in Equation (4), \( \mu'_3 \) is expressed in terms of \( \tau_1 \) and \( \tau_1 \). In order to express \( \mu'_3 \) in terms of \( \tau_1 \) and \( \tau_1 \), the software Mathematica\(^14\) is used. The final form of \( \mu'_3 \) is as follows
Studying the Third Cumulant of the Mixture of Dirichlet-multinomial Distributions

\[\begin{align*}
\mu^3_3 &= m(p_{1j}w_1 + p_{2j}w_2)(1 - p_{1j}w_1 - p_{2j}w_2)(1 - 2p_{1j}w_1 - 2p_{2j}w_2), \\
\kappa_3 &= m^3 - 3m^2\mu^3_3 + 2m\mu^3_3. \\
\kappa_3^0 &= m\left(p_{1j}w_1 + p_{2j}w_2\right)\left(1 - p_{1j}w_1 - p_{2j}w_2\right)(1 - 2p_{1j}w_1 - 2p_{2j}w_2).
\end{align*}\]

Now using Equations (1), (3) and (5) the 3\textsuperscript{rd} cumulant of the mixture of Dirichlet-multinomial distribution can be calculated as

\[\begin{align*}
\kappa_3 &= \mu^3_3 - 3m\mu^2_3 + 2m\mu^3_3 \\
&= 2(m p_{1j}w_1 + m p_{2j}w_2)^3 \\
&\quad + m \left(p_{1j}w_1 - 3m - 1)p_{1j}(p_{1j}(r_1 - 1) - r_1)w_1^2\right) \\
&\quad + \left(m + 2\right)(m - 1)p_{1j}(p_{1j}(r_1 - 1) - r_1)w_1^2 \\
&\quad + m p_{2j}w_2 \\
&\quad - 3m p_{1j}p_{2j}w_1 w_2 \left(\frac{m - 1}{1 + r_1} - 1\right) - 3m(p_{1j}w_1 + p_{2j}w_2) \\
&\quad - 3m p_{1j}p_{2j}w_1 w_2 \left(\frac{m - 1}{r_2 - 1} - 1\right) - 3m(p_{1j}w_1 + p_{2j}w_2) \\
&\quad - 3m p_{1j}p_{2j}w_1 w_2 \left(\frac{m - 1}{r_2 - 1} - 1\right) - 3m(p_{1j}w_1 + p_{2j}w_2) \\
&\quad - 3m p_{1j}p_{2j}w_1 w_2 \left(\frac{m - 1}{r_2 - 1} - 1\right) - 3m(p_{1j}w_1 + p_{2j}w_2).
\end{align*}\]  

Now, the 3\textsuperscript{rd} cumulant of multinomial distribution is

\[\kappa_3^0 = m\left(p_{1j}w_1 + p_{2j}w_2\right)\left(1 - p_{1j}w_1 - p_{2j}w_2\right)(1 - 2p_{1j}w_1 - 2p_{2j}w_2).\]  

From Afroz\textsuperscript{10} the third cumulant should satisfy the assumption \(\kappa_3 = \alpha\kappa_3^0\), where \(\alpha \geq \varphi^2\), i.e. \(\alpha\) must be positive. Here using Equations (6) and (7) one can calculate the value of \(\alpha = \frac{\kappa_3}{\kappa_3^0}\). Since, the form of \(\alpha\) would be very complex, \(\alpha\) is plotted for different values of the parameters \((m, p_1, p_2, w_1, w_2, r_1, r_2)\) using Mathematica\textsuperscript{14}. 

V. Discussion and Conclusion

In this paper, the 1\textsuperscript{st} to 3\textsuperscript{rd} order raw moments and the 3\textsuperscript{rd} cumulant \((\kappa_3)\) of the mixture of Dirichlet-multinomial distributions are derived. Higher order moments are useful to develop goodness of fit statistics, also use of additional information about the moments in estimation procedure can provide more efficient estimators. Here it is found that, \(\kappa_3\) does not satisfy the assumption discussed in Fletcher\textsuperscript{9} and Afroz\textsuperscript{10}. Unlike the present set up one can consider other type of mixtures of Dirichlet-multinomial distributions, where one subpopulation is selected randomly from the two sub-populations and the whole
sample comes from the selected sub-population. Also, there are other kind of distributions to model overdispersed multinomial data such as finite-mixture distribution (Morel and Nagaraj). The future research may involve derivation of the moments of such distributions and also checking the assumption on the 3rd cumulant. Furthermore, a new estimator of overdispersion which is more relaxed to the assumption on the third cumulant can be developed.

References