Dealing with overdispersion in multivariate count data

Dealing with Overdispersion in Multivariate Count Data

Noemi Corsini, Cinzia Viroli

Department of Statistical Sciences, University of Bologna
via Belle Arti 41, 40126, Bologna, Italy

Abstract

The problem of overdispersion in multivariate count data is a challenging issue. It covers a central role mainly due to the relevance of modern technology-based data, such as Next Generation Sequencing and textual data from the web or digital collections. A comprehensive analysis of the likelihood-based models for extra-variation data is presented. Particular attention is paid to the models feasible for high-dimensional data. A new approach together with its parametric-estimation procedure is proposed. It can be viewed as a deeper version of the Dirichlet-Multinomial distribution and it leads to important results allowing to get a better approximation of the observed variability. A significative comparison of the proposed model and existing strategies is made through two different simulation studies and an empirical data set, that confirm a better capability to describe overdispersion.

Keywords: Extra-variation, Mixture models, Deep Learning, Maximum Likelihood

1. Introduction

The overdispersion or extra-variation is a recurring phenomenon when dealing with counts and categorical data. In particular, it often occurs that after fitting a binomial, a multinomial or a Poisson model to the data, the sampling

Email address: cinzia.viroli@unibo.it (Cinzia Viroli)

variation is greater than the estimated variation accounted by the model (see for instance [1]). In other words, the data exhibit a larger variability than that the model is able to explain [2]. Overdispersion has specific causes and consequences. It may arise as result of the data collection and aggregation, such as clumped sampling [3] or it may due to correlation between individual responses or to additional experimental variability. Inferential consequences are imprecise estimates and biased standard errors that make model selection, interpretability and prediction unreliable.

We focus our analysis on multivariate count data, that are becoming more recurrent thanks to recent technologies such as web scraping for textual data [4] or Next Generation Sequencing data [5]. In both situations, we observe multivariate count data often inflated by a large amount of zeros (words rarely used or not-expressed genes) or correlated responses. As a consequence, extra-variation is typically observed, and the phenomenon is particularly reinforced by the limited number of replicates and high-dimensionality.

The multinomial distribution is the natural probabilistic model to describe multivariate count data but, in presence of overdispersion, it typically leads to nominal variances well below to the empirical variability. It is possible to cope with overdispersion by several strategies.

Zero-inflated probabilistic models are one of the most common strategies to deal with extra-variation due to zeros for univariate variables. Among these, zero-inflated Poisson and zero-inflated Binomial distribution proved to be very efficient for count data with excess of zeros (see, for instance, [6, 7]). However, for multivariate count data, the zero-inflated multinomial distribution assumes that there are specific zero-inflated categories and at least one non-inflated category common to all observations that are known from the empirical context [8]. This strong assumption does not make the approach applicable to general situations with arbitrary sparsity or simply extra-variation not due to zeros.

Quasi-likelihood assumes that the variance depends on a dispersion parameter, say ρ, representing overdispersion [9] and instead of defining a probabilistic form for the distribution of the data it is sufficient to specify only the variance-

mean relationship. A specific quasi-likelihood approach for multivariate count data was investigated by [10]. In the recent years, some generalizations have been proposed: [11] developed an alternative way of estimating ρ when data are sparse, while [12] explained how to deal with clustered multinomial data and unequal cluster size.

Despite the quasi-likelihood approach is robust and works well with severe overdispersion, the problem can be also dealt with alternative and extended family of distributions in a maximum-likelihood perspective. Among these, the Dirichlet-Multinomial compound model [13] represents one of the most common solutions, able to capture extra-variability by a simple prior on the multinomial parameters.

The study and the comparison of the main probabilistic models of the statistical literature able to capture extra-variation in multivariate count data are our focus. A new model that extends the Dirichlet-Multinomial in a deep fashion is also presented together with its parametric-estimation procedure. More precisely, the model resembles the deep learning architecture composed by an additional hidden layer with several nodes [14]. A relevant aspect of this model is that its variance tends to the computed variance when the number of nodes goes to $+\infty$, as empirically shown in the simulation study.

The paper is organized in the following way. In Section 2 we present the main parametric models of the statistical literature accounting for extra-variation. We will examine in depth the approaches that are adequate to deal with high-dimensional data. The proposed strategy and its estimation procedure are introduced in Section 3. A simulation study showing the empirical performance of the different strategies is presented in Section 4, while an empirical application to RNA sequencing data is developed in Section 5. Conclusions and final remarks can be found in Section 6.

2. Models for overdispersion

Let $\mathbf{Y} = (Y_1, \ldots, Y_j, \ldots, Y_p)$ be a multivariate vector of counts, where p denotes the total number of categories. In the Multinomial distribution

$$P(\mathbf{Y} = \mathbf{y}) = \frac{m!}{y_1! \ldots y_p!} \prod_{j=1}^{p} \pi_j^{y_j} \tag{1}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)^T$ represents the success probability of each of the p categories with $0 \leq \pi_j \leq 1$ and $\sum_{j=1}^{p} \pi_j = 1$ and $m = \sum_{j=1}^{p} y_j$ is the size indicating the total number of independent trials. The mean and the variance of the distribution depend on $\boldsymbol{\pi}$ and m through

$$E[\mathbf{Y}] = m\boldsymbol{\pi} \tag{2}$$

$$Var[\mathbf{Y}] = m\{diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T\}. \tag{3}$$

The multinomial distribution naturally describes the outcomes of m independent trials into p categories, but in many practical situations the assumption of independence of the trials is not respected resulting in the phenomenon of extra-multinomial variation, as shown in [13]. Another aspect of this distribution is that it models negative correlations between categories, as clear by taking the marginals in (3) that are $Covar[Y_j, Y_{j'}] = -m\pi_j\pi_{j'}$.

Alternative parametric extra-variation models have been proposed in the literature; they may be distinguished by the reason behind the lack of independence.

Dirichlet-Multinomial. The first parametric alternative to the multinomial distribution was derived by [13], under the assumption that the multinomial probability parameters $\pi_1, ..., \pi_p$ are distributed according to a Dirichlet distribution. Since it is the natural conjugate of the multinomial, the resulting compound distribution has a closed form and it takes the name of Dirichlet-Multinomial (DM). It is also known in the statistical literature as Multivariate Pólya distribution, it being the multivariate version of the Beta-Binomial distribution.

From the compound of the Dirichlet distribution with the Multinomial, the random vector $\mathbf{Y} \sim DM_p(\boldsymbol{\theta}, m)$ has probability function:

$$P(\mathbf{Y} = \mathbf{y}) = \frac{\Gamma(\theta_0)\Gamma(m+1)}{\Gamma(m+\theta_0)} \prod_{j=1}^{p} \frac{\Gamma(y_j+\theta_j)}{\Gamma(\theta_j)\Gamma(y_j+1)}$$

where $\theta_0 = \sum_{j=1}^{p} \theta_j$ and Γ is the gamma function. By denoting with $\boldsymbol{\pi} = (\frac{\theta_1}{\theta_0}, \ldots, \frac{\theta_p}{\theta_0})$ it is possible to show that the expectation is

$$E[\mathbf{Y}] \quad = \quad m\boldsymbol{\pi} \tag{4}$$

so that it has the same expression of the Multinomial expectation in (2). The resulting variance is corrected by a term in order to account for the extra-variation of the data

$$Var[\boldsymbol{Y}] = m\{1+\rho^2(m-1)\}\{diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'\} \tag{5}$$

where ρ is the overdispersion parameter defined through $\rho^2 = \frac{1}{1+\theta_0}$, so that $0 < \rho < 1$. The constant $1+\rho^2(m-1)$ inflates the variance of the multinomial distribution and this is what makes the DM a good distribution for modeling overdispersion. Notice that when $\rho = 0$ the DM distribution coincides with the multinomial one. Having the same kernel form of the multinomial distribution, it is easy to check that the correlations among variables are negative.

A recent extension of the DM has been proposed by [15]; in this framework the multivariate beta distribution [16] is proposed as prior, resulting in a very flexible model. The model is estimated via an independent Metropolis-Hastings algorithm that makes the fitting computationally demanding as the number of replicates and categories increase.

Random-Clumped Multinomial. The Random-Clumped Multinomial (RCM) was proposed by [17] as an alternative to the Dirichlet-Multinomial distribution with the idea to describe the extra-multinomial variation introduced by correlation or clumped multinomial sampling. Specifically, clumped sampling refers to a sample where the values observed in correspondence of each statistical unit are influenced by the value of others, i.e. the statistical units are not independent.

In RCM the vector of counts $\mathbf{Y}$ originates by two parts: the first one takes into account the possibility that in cluster sampling within the cluster there are some identical responses due to individuals that greatly influence each other; the second part considers the remaining independent responses. Formally:

$$\boldsymbol{Y} = \boldsymbol{X}N + (\boldsymbol{Z} \mid N) \tag{6}$$

where $\mathbf{X}$ is distributed as a multinomial with size 1 and p categories, say $M_p(\boldsymbol{\pi}, 1)$, independently from $N \sim M_2(\rho, m)$, which has a binomial distribution. In the second term, $(\boldsymbol{Z} \mid N) \sim M_p(\boldsymbol{\pi}, m - N)$ if $N < m$. The random number of counts N is added to $\mathbf{X}$ meaning that the addend $\mathbf{X}N$ replicates N times the response given by $\mathbf{X}$, whereas $(\boldsymbol{Z} \mid N)$ considers the independent responses.

It is possible to prove that the probability distribution of $\mathbf{Y}$ is a finite mixture of multinomials [17] and more precisely:

$$P(\boldsymbol{Y} = \boldsymbol{y}) = \sum_{j=1}^{p} \pi_j P(W_j = \boldsymbol{y}) \tag{7}$$

where W_j for $j = 1, \ldots, p-1$ is distributed according to a $M_p((1-\rho)\boldsymbol{\pi} + \rho e_j; m)$ and $W_p \sim M_p((1-\rho)\boldsymbol{\pi}; m)$, e_j is the j-th column of the $(p-1) \times (p-1)$ identity matrix and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)'$ is a probability vector used both as weights for the mixture and as parameters of the multinomials considered in the mixture itself. The RCM distribution has the same mean and variance of the DM distribution, therefore, theoretically speaking, it can describe the same amount of extra-variation. Empirical differences are thus only ascribed to the estimation method.

The original model proposed by [17] accounts for a single clumping only; [18] developed an extension which integrates multiple random clumping. Specifically they show that the extended finite mixture distribution is a multinomial mixing distribution with different mixing coefficients. This allows to introduce more flexibility but at the cost of additional complexity.

The model can be estimated through the Fisher's scoring method. [19] proposed a two-stage procedure in computing the maximum likelihood estimates in

which at first the algorithm uses theoretical limiting results until convergence and then in the second step an extra iteration with the exact Fisher information matrix is implemented. The resulting algorithm is less computationally expensive with respect to a simple Fisher scoring algorithm, and, at the same time, leads to a better accuracy. As alternative solution, [20] developed a very fast estimation procedure based on an hybrid approach. At first an approximation of the Fisher scoring algorithm is considered; after an initial warm-up, the classical Fisher's scoring algorithm is applied. More recently, a minorization-maximization algorithm for fitting the RCM has been proposed by [21].

Negative Multinomial. In the multinomial distribution it is well known that the marginals are binomial variates exhibiting a negative correlation. The same negative association between variables is inherited by the DM and the RCM distributions. The Negative Multinomial (NM) distribution assumes instead a positive correlation between variables. It is simply a generalization of the Negative Binomial when multiple outcomes are considered [22]. In the NM, $\boldsymbol{Y}$ has parameters $(\boldsymbol{\pi}, \beta) = (\pi_1, \ldots, \pi_{p+1}, \beta)$, $\sum_{j=1}^{p+1} \pi_j = 1$, $\beta > 0$ and the probability mass function is defined as

$$P(\boldsymbol{Y} = \boldsymbol{y}) = \binom{\beta + m - 1}{m} \binom{m}{\boldsymbol{y}} \prod_{j=1}^{p} \pi_j^{y_j} \pi_{p+1}^{\beta}, \tag{8}$$

where m is the size and $\pi_{p+1} = 1 - \sum_{j=1}^{p} \pi_j$ is the probability of a failure. The first two moments of this distribution are the following:

$$E[\mathbf{Y}] = \beta \frac{\boldsymbol{\pi}}{\pi_{p+1}}, \tag{9}$$

$$Var[\mathbf{Y}] = \frac{\beta}{\pi_{p+1}^2} \boldsymbol{\pi}\boldsymbol{\pi}' + \frac{\beta}{\pi_{p+1}} diag(\boldsymbol{\pi}). \tag{10}$$

The model can be fitted via maximum likelihood by an iteratively reweighted Poisson regression (see [22] and [23] for further details).

Generalized Dirichlet Multinomial. The Generalized Dirichlet-Multinomial (GDM) was proposed by [24] with the aim to have a general covariance matrix and correlation structure among variables. The basic idea is to choose a more flexible

mixing distribution as a prior for the multinomial given by a kind of generalized Dirichlet distribution. Following the notation of [22], the probability mass function of the GDM is

$$P(\boldsymbol{Y}=\boldsymbol{y}) = \frac{m!}{y_1!\dots y_p!}\prod_{j=1}^{p-1}\frac{\Gamma(\alpha_j+y_j)}{\Gamma(\alpha_j)}\frac{\Gamma(\beta_j+\sum_{h=j}^{k}y_h)}{\Gamma(\beta_j)}\frac{\Gamma(\alpha_j+\beta_j)}{\Gamma(\alpha_j+\beta_j+\sum_{h=j}^{k}y_h)} \quad (11)$$

where $(\boldsymbol{\alpha},\boldsymbol{\beta}) = (\alpha_1,\dots,\alpha_{p-1},\beta_1,\dots,\beta_{p-1})$ are the parameters of this distribution, with $\alpha_j,\beta_j > 0$. When $\beta_j = \sum_{h=j+1}^{p}\alpha_h$ the GDM reduces to the DM distribution.

The distribution has the following expectation and variance

$$E[Y_j] = m\begin{cases}\dfrac{\alpha_1}{\alpha_1+\beta_1} & j=1\\ \dfrac{\alpha_j}{\alpha_j+\beta_j}\displaystyle\prod_{h=1}^{j-1}\frac{\beta_h}{\alpha_h+\beta_h} & j=2,...,p-1\\ \displaystyle\prod_{j=1}^{p-1}\frac{\beta_j}{\alpha_j+\beta_j} & j=p\end{cases} \quad (12)$$

$$\begin{aligned}Var[Y_j] = {} & m\frac{\alpha_j}{\alpha_j+\beta_j}\prod_{h=1}^{j-1}\frac{\beta_h}{\alpha_h+\beta_h}\left[(m-1)\prod_{h=1}^{j-1}\frac{\beta_h+1}{\alpha_h+\beta_h+1}\frac{\alpha_j+1}{\alpha_j+\beta_j+1}\right.\\ & \left.-m\prod_{h=1}^{j-1}\frac{\beta_h}{\alpha_h+\beta_h}\frac{\alpha_j}{\alpha_j+\beta_j}+1\right]\end{aligned} \quad (13)$$

Thanks to the generalized prior of the GDM distribution it is possible to get both positive and negative pairwise correlations between the marginals. The estimation of this model can be obtained via maximum likelihood with quasi-Newton iterations (see [23] for major details).

In addition to these important contributions, other models and extensions were introduced over time to deal with overdispersion. Interesting recent works focused on Conway-Maxwell-Multinomial [25] and Multiplicative Multinomial model [26]. Both strategies are very flexible and they allow for both overdispersion and underdispersion but they incur in a heavy computational burden, making them infeasible for high-dimensional data, like the ones considered in Section 5 .

3. Deep Dirichlet-Multinomial

3.1. Model definition

In order to deal with overdispersion, in this section we propose a new model that consists of a special kind of a mixture of Dirichlet-Multinomial distributions with restrictions on the parameters. This model, called Deep Dirichlet-Multinomial (DDM), is derived from the mixture model developed by [27].

More precisely, let $DM(\boldsymbol{\theta}, m)$ be the probability mass function of a Dirichlet-Multinomial with parameters $\boldsymbol{\theta}$ and size m, then the probability distribution of the DDM model is defined as

$$P(\boldsymbol{Y} = \boldsymbol{y}) = \sum_{k=1}^{K} w_k DM(\boldsymbol{\beta}(1 + \boldsymbol{\alpha_k}), m), \tag{14}$$

where w_k for $k = 1, \ldots, K$ represents the generic element of the vector of weights $\mathbf{w} = (w_1, w_2, \ldots, w_{K-1})$ with $w_K = 1 - \sum_{k=1}^{K-1} w_k$ and $0 < w_k < 1$. A graphical representation of the DDM structure is shown in Figure 1. As clear from the depicted structure, a hidden layer of nodes is introduced to better capture the overdispersion. The adjective 'deep' highlights the flexibility of the introduced hidden mixture.

In the model, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha_k}$ are vectors of length p with components satisfying $\beta_j > 0$ and $-1 < \alpha_{jk} < 1$ $(j = 1, \ldots, p)$; thus each $\boldsymbol{\alpha}_k$ can be interpreted as a perturbation parameter. It being defined in (-1,1), its role is to adjust $\boldsymbol{\beta}$ and to get a more flexible model that behaves better in case of overdispersion. Positive values of α_j lead to larger effects of $\theta_j = \beta_j(1 + \alpha_j)$ and negative values of α_j lead to lower $\theta_j = \beta_j(1 + \alpha_j)$, thus indicating which categories have more zeros. Figure 2 shows the effect of the parameters $\boldsymbol{\alpha_k}$ on the estimated $\boldsymbol{\beta}$ on simulated data (n=200 and p=10) for a DMM with $K = 2$. The solid line represents the values of $\boldsymbol{\beta}$ in increasing order and the dashed lines depict the perturbed parameters by the effect of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$. As clear from the graph, the two perturbation parameters tend to increase and decrease the not-perturbed $\boldsymbol{\beta}$, respectively.

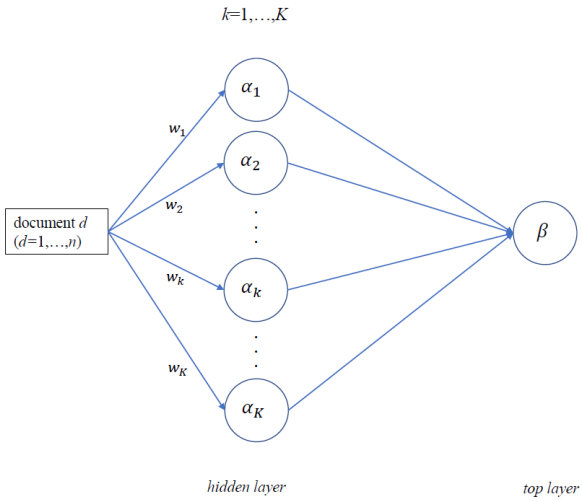

Figure 1: Structure of the DDM model.

The expectation of the distribution is the weighted sum of the expected values of each DM distributions:

$$E[\boldsymbol{Y}] = \sum_{k=1}^{K} w_k m \boldsymbol{\pi}_k. \tag{15}$$

In order to derive the variance of the model, we use the moment generating function. Let $\boldsymbol{\theta}_k = \boldsymbol{\beta}(1 + \boldsymbol{\alpha_k})$, $\theta_{0k} = \sum_{j=1}^{p} \beta_j(1 + \alpha_{jk})$ and $\boldsymbol{\pi}_k = \frac{\boldsymbol{\theta}_k}{\theta_{0k}}$, then the moment generating function of the mixture is

$$\begin{aligned} \phi_Y(t) &= \sum_{k=1}^{K} w_k \phi_{x_k}(t) \\ &= \sum_{k=1}^{K} w_k \frac{\Gamma(m+1)\Gamma(\theta_{0k})}{\Gamma(m+\theta_{0k})} D_m(\theta_k, (e^{t_1}, \ldots, e^{t_p})) \\ \text{with } D_m &= \frac{1}{m} \sum_{u=1}^{m} \left[\left(\sum_{j=1}^{p} \theta_{jk} e^{t_j u} \right) D_{m-u} \right], D_0 = 1, \end{aligned} \tag{16}$$

where specifically $\phi_{x_k}(t)$ is the moment generating function of a $DM(\boldsymbol{\theta}_k, m)$. By using the previous moment generating function we can derive the second moment and the variance of the DDM distribution. It is not difficult to prove

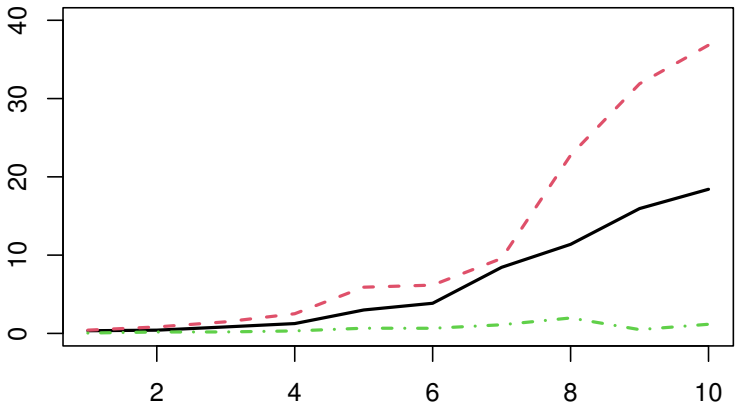

Figure 2: Effect of the perturbation parameters $\boldsymbol{\alpha}$ on $\boldsymbol{\beta}$, with $p = 10$. The solid black line represents the p values of the original β_j in increasing order, the dashed lines represent the perturbed parameters.

that the variance can be split into two components, the first one is a weighted sum of within variances, the second term is a sort of between variance part. Formally:

$$\begin{aligned} Var[\boldsymbol{Y}] &= \sum_{k=1}^{K} w_k m\{diag(\boldsymbol{\pi}_k) - \boldsymbol{\pi}_k\boldsymbol{\pi}_k'\}(1+\rho_k^2(m-1)) \\ &+ \sum_{k=1}^{K} w_k m^2 \boldsymbol{\pi}_k\boldsymbol{\pi}_k' - m^2\left(\sum_{k=1}^{K} w_k\boldsymbol{\pi}_k\right)\left(\sum_{k=1}^{K} w_k\boldsymbol{\pi}_k\right)', \end{aligned} \tag{17}$$

where $\rho_k^2 = 1/(1+\theta_{k0})$. The between variance is an additional addendum that can capture both over- and under-dispersion. By marginalizing the quantity along two different categories, say j and j', the covariance formula is straight-

forward:

$$\begin{aligned} Covar[Y_j, Y_{j'}] &= \sum_{k=1}^{K} w_k \pi_{jk} \pi_{j'k} (1 - \rho_k^2) m(m-1) \\ &- m^2 \left(\sum_{k=1}^{K} w_k \pi_{jk} \right) \left(\sum_{k=1}^{K} w_k \pi_{j'k} \right). \end{aligned} \tag{18}$$

The expression can take both positive and negative values denoting that the distribution is able to cope with flexible correlation structures among variables. This is a very important property in practice, since groups of variables could be positively correlated to each other but negatively correlated with other variables. For instance, genes could be co-expressed together or, alternatively, words used together in the same context, but synonymous are negatively correlated. Here, this extreme flexibility is obtained at the price of many parameters to be estimated, which largely increase with K.

Another important result that we will show empirically is that the variance of this model tends to the empirical variance when the number of components of the mixture K goes to $+\infty$. This is strictly related to the property that mixture models are universal approximator of densities [28, 29]. In other terms, given any probability density distribution, there exists a mixture model (with possibly many components) such that the distribution of the mixture approximates the given distribution with arbitrary precision. Thus we expect that proposed deep Dirichlet-Multinomial distribution will capture extra-variation provided that K increases. Notice that the Random-Clumped Multinomial shares the property to be a particular mixture model. However, it is restricted to have fixed components equal to the number of categories p, thus making it attractive as approximator when p increases. In subsection 4.3 the issue will be empirically analyzed through two simulation studies.

3.2. Model estimation

Given a set of observations $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ under the assumption of IID random variables, the log-likelihood of the model can be written as

$$\begin{aligned} \ell(\boldsymbol{\Theta}) &= \sum_{i=1}^{n} \log \sum_{k=1}^{K} w_k DM(\boldsymbol{\beta}(1+\boldsymbol{\alpha}_k), m) \\ &= \sum_{i=1}^{n} \log \sum_{k=1}^{K} w_k \frac{\Gamma(\theta_{0k})\Gamma(m+1)}{\Gamma(\theta_{0k}+m)} \prod_{j=1}^{p} \frac{\Gamma(y_{ij}+\theta_{jk})}{\Gamma(\theta_{jk})\Gamma(y_{ij}+1)} \end{aligned} \tag{19}$$

where $\boldsymbol{\Theta}$ denotes the full set of parameters, and, as defined before, $\boldsymbol{\theta}_k = \boldsymbol{\beta}(1+\boldsymbol{\alpha}_k)$ and $\theta_{0k} = \sum_{j=1}^{p} \theta_{jk}$.

Parameters in (19) can be efficiently estimated through a generalized EM algorithm [30] with a quasi-Newton optimization step for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}_k$. The EM algorithm maximizes the conditional expectation of the so-called complete density given the observable data and alternates between the expectation and the maximization steps until convergence. Let z be the allocation variable of the mixture model defined in (14) denoting the component membership of each observation. By definition z follows a multinomial distribution

$$f(z|\boldsymbol{\Theta}) = \prod_{k=1}^{K} w_k^{z_k},$$

from which $f(z_k = 1|\boldsymbol{\Theta}) = w_k$. Evidently, the conditional density of each $\boldsymbol{y}_i$, given the allocation variable, is the kth DM distribution.

Then the parameter function to be maximized is the conditional expectation of the complete density $f(\boldsymbol{y}, z|\boldsymbol{\Theta})$ given the observable data, using a fixed set of parameters $\boldsymbol{\Theta}'$:

$$\begin{aligned} & \arg\max_{\boldsymbol{\Theta}} E_{z|\boldsymbol{y};\boldsymbol{\Theta}'} \left[\log f\left(\boldsymbol{y}, z|\boldsymbol{\Theta}\right)\right] \\ = \; & \arg\max_{\boldsymbol{\Theta}} E_{z|\boldsymbol{y};\boldsymbol{\Theta}'} \left[\log f\left(\boldsymbol{y}|z;\boldsymbol{\Theta}\right) + \log f\left(z|\boldsymbol{\Theta}\right)\right]. \end{aligned} \tag{20}$$

By observing $f(\boldsymbol{y}|z, \boldsymbol{\Theta}) = \prod_{k=1}^{K} DM(\boldsymbol{y}_i; \boldsymbol{\beta}(1+\boldsymbol{\alpha}_k), m)^{z_k}$ it is easy to see that formula (20) is equivalent to maximizing the following function with respect to

$\boldsymbol{\Theta}$:

$$\begin{aligned} L\left(\boldsymbol{\Theta}\right) &= \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}\log\left[w_k DM(\boldsymbol{y}_i;\boldsymbol{\beta}(1+\boldsymbol{\alpha}_k),m)\right] \\ &\quad \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}\log w_k + \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}\log DM(\boldsymbol{y}_i;\boldsymbol{\beta}(1+\boldsymbol{\alpha}_k),m) \quad (21) \end{aligned}$$

where τ_{ik} is the posterior probability that $\boldsymbol{y}_i$ belongs to the kth component of the mixture:

$$\tau_{ik} = \frac{w_k DM(\boldsymbol{y}_i;\boldsymbol{\beta}(1+\boldsymbol{\alpha}_k),m)}{\sum_{h=1}^{K} w_h DM(\boldsymbol{y}_i;\boldsymbol{\beta}(1+\boldsymbol{\alpha}_h),m)}. \quad (22)$$

At each iteration, in the E-step we compute the posterior distributions τ_{ik} as function of the current set of parameters. In the M-step we separately maximize the two terms in (21) under the parameter constraints.

The estimation of the mixture weights under the constraints that they are positive and sum to one takes the closed-form formula:

$$\hat{w}_k = \frac{\sum_{i=1}^{n}\tau_{ik}}{n}. \quad (23)$$

Maximization of the positive vectors $\boldsymbol{\beta}$ and constraint vectors $\boldsymbol{\alpha}_k$ involves the derivative of $\log P\left(\boldsymbol{y}_i|z_i=k;\boldsymbol{\Theta}\right)$ that can be rewritten as

$$\begin{aligned} \log P(\boldsymbol{y}_i|z_i=k;\boldsymbol{\Theta}) &\propto \log\Gamma\left(\sum_{j=1}^{p}\theta_{jk}\right) - \log\Gamma\left(\sum_{j=1}^{p}y_{ij}+\theta_{jk}\right) \\ &- \sum_{j=1}^{p}\log\Gamma\left(\theta_{jk}\right) + \sum_{j=1}^{p}\log\Gamma\left(y_{ij}+\theta_{jk}\right). \end{aligned}$$

By remembering $\theta_{jk} = \beta_j(1+\alpha_{jk})$, the gradient of the previous term with respect to the vectors $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}_k$ can be obtained as function of digamma defined as $\psi(x) = \frac{d}{dx}\log\Gamma(x)$. Let $\mathbf{1}$ be a column vector of ones of length p. The score with respect to $\boldsymbol{\beta}$ is

$$\begin{aligned} &\frac{\partial \log P(\boldsymbol{y}_i|z_i=k;\boldsymbol{\Theta})}{\partial\boldsymbol{\beta}} = S_k(\boldsymbol{\beta}) = \psi\left(\boldsymbol{\theta}_k^\top\mathbf{1}\right)(1+\boldsymbol{\alpha}_k^\top) \\ &-\psi\left(\sum_{j=1}^{p}y_{ij}+\theta_{jk}\right)(1+\boldsymbol{\alpha}_k^\top) - (\psi(\theta_{1k})(1+\alpha_{1k}),\ldots,\psi(\theta_{pk})(1+\alpha_{pk})) \\ &+(\psi(\theta_{1k}+y_{i1})(1+\alpha_{1k}),\ldots,\psi(\theta_{pk}+y_{ip})(1+\alpha_{pk})). \end{aligned}$$

Similarly, the score with respect to $\boldsymbol{\alpha}_k$ is

$$\frac{\partial \log P(\boldsymbol{y}_i|z_i = k; \boldsymbol{\Theta})}{\partial \boldsymbol{\alpha}_k} = S_k(\boldsymbol{\alpha}_k) = \psi\left(\boldsymbol{\theta}_k^\top \mathbf{1}\right)\boldsymbol{\beta}_k^\top$$
$$-\psi\left(\sum_{j=1}^{p} y_{ij} + \theta_{jk}\right)\boldsymbol{\beta}^\top - (\psi(\theta_{1k})\beta_1, \ldots, \psi(\theta_{pk})\beta_p)$$
$$+ (\psi(\theta_{1k} + y_{i1})\beta_1, \ldots, \psi(\theta_{pk} + y_{ip})\beta_p)\,.$$

Given these scores it is evident that no solution exists in closed form. However at each iteration of the EM algorithm, estimates can be obtained according to quasi-Newton strategies. The scheme of the algorithm is the following:

---

1. *Initialization*: Set $h = 0$. For each component $k = 1, \ldots, K$, choose values for the vectors $\boldsymbol{\alpha}_k^{(h)}$ and $\boldsymbol{\beta}^{(h)}$ and fix equispaced probabilities for $w_k^{(h)}$.
2. *Estimation step*: Repeat the following until $\ell(\boldsymbol{\Theta})$ stops changing:
   (a) Compute the posteriors using (22);
   (b) For $k = 1, \ldots, K$ compute new values for $\boldsymbol{\alpha}_k$ using the scores $S_k(\boldsymbol{\alpha}_k)$ by constrained quasi-Newton.
   (c) Compute new values for $\boldsymbol{\beta}$ using the weighted sum of scores $\sum_{k=1}^{K} w_k S_k(\boldsymbol{\beta})$ by constrained quasi-Newton.
   (d) For $k = 1, \ldots, K$ compute new values for w_k using (23).
   (e) $h = h + 1$.

The algorithm has been implemented in R code and is available via Github[1].

4. Empirical results

Table 1 contains a synthetic summary of the main characteristics of the presented distributions for multivariate count data.

[1] `https://github.com/NoeCors/EM_Mixtures_DM/tree/main`

	Multinomial (MN)	*Dirichlet-Multinomial (DM)*
Parameters	$m,\ \boldsymbol{\pi} = (\pi_1, \ldots, \pi_{p-1})'$	$m,\ \rho,\ \boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$
	$\sum_{j=1}^{p} \pi_j = 1$	$\theta_0 = \sum_{j=1}^{p} \theta_j$, $\theta^2 = \frac{1}{1+\theta_0}$, $\boldsymbol{\pi} = \frac{\boldsymbol{\theta}}{\theta_0}$
♯ parameters	p	$p+1$
Expectation	$m\boldsymbol{\pi}$	$m\boldsymbol{\pi}$
Variance	$m\{diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T\}$	$m\{1+\rho^2(m-1)\}\{diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'\}$
Covariance	Negatively correlated	Negatively correlated
	Random-Clumped Mult. (RCM)	*Negative Multinomial*
Parameters	$m,\ \rho, \boldsymbol{\pi} = (\pi_1, \ldots, \pi_{p-1})'$	$m,\ \beta, \boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)'$
	$0 < \rho < 1$ and $\sum_{j=1}^{p} \pi_j = 1$	$\beta > 0$ and $\pi_{p+1} = 1 - \sum_{j=1}^{p} \pi_j$
♯ parameters	$p+1$	$p+2$
Expectation	$m\boldsymbol{\pi}$	$\beta\frac{\boldsymbol{\pi}}{\pi_{p+1}}$
Variance	$m\{1+\rho^2(m-1)\}\{diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'\}$	$\frac{\beta}{\pi_{p+1}^2}\boldsymbol{\pi}\boldsymbol{\pi}' + \frac{\beta}{\pi_{p+1}}diag(\boldsymbol{\pi})$
Covariance	Negatively correlated	Positively correlated
	Generalized DM (DGM)	*Deep DM (DDM)*
Parameters	$m,\ \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{p-1})$	$m,\ (w_1, \ldots, w_{K-1}),\ w_K = 1 - \sum_{k=1}^{K-1}$
	$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p-1})$	$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p),\ \boldsymbol{\alpha}_k = (\alpha_1, \ldots, \alpha_p)$
	$\alpha_j > 0,\ \beta_j > 0$	$w_k > 0,\ \beta_j > 0,\ -1 < \alpha_j < 1$
♯ parameters	$2p-1$	$p(K+1)+K$
Expectation	see equation (12)	$\sum_{k=1}^{K} w_k m\boldsymbol{\pi}_k$
Variance	see equation (13)	see equation (17)
Covariance	General correlation	General correlation

Table 1: Models for multivariate count data

In this section we investigate the capability of the proposed model together with the existing illustrated strategies to capture overdispersion through two simulation studies.

4.1. Performance comparison

We consider two empirical studies differing in the way the overdispersion is introduced.

More specifically, an increasing percentage of zeros is introduced into the data in order to gradually check the capability of the different probabilistic models to deal with overdispersion. Data are first randomly generated by a multinomial distribution. Then, in the first scenario, the zeros are added in

a completely random way into the dataset. In the second scenario, we added the zeros by gradually replacing the smallest counts, starting from cells with frequency one, ending up to larger counts.

We make a comprehensive comparison of the likelihood-based models discussed up until now and summarized in Table 1. To this aim, we generated 100 datasets with 10 levels of increasing overdispersion. The ten levels correspond to an increasing proportion of zeros, through jumps of 10%, starting from the case of lack of extra-variation with a percentage of added zeros equal to zero to the case of maximum overdispersion of the data with a percentage of added zeros equal to 90%.

The models are fitted on each dataset of the two empirical studies considering datasets with different combinations of samples n and categories p randomly generated from a multinomial distribution with parameters $m = 100$ and $\boldsymbol{\pi} \sim Unif[0, 1]$ then normalized. Here we present the results for $n = 200$ rows and $p = 10$ columns. With respect to the DDM distribution, we report four cases each one differentiated by the number of mixture components $K = 2$, $K = 3$, $K = 4$ and $K = 20$.

In Table 2 the analytical results of the comparison of the different methods are shown. In particular, the table displays the average Euclidean distances between the estimated variances of each model and the empirical ones, the average BIC [31] and average AIC [32] across the replicated datasets and the different settings of added zeros. The last column shows the average of the computational times (in seconds) to get estimates on the datasets. In general, flexibility comes at the price of a greater computational burden. The DDM model with 20 mixture components is the most demanding estimation method from the computational time perspective, but times remain generally feasible.

According to these results, it is clear that the proposed Deep Dirichlet-Multinomial is the model able to better describe the variability of the data, it having the smallest euclidean distance. However, this is achieved at the price of a large number of parameters. In fact, the two information criteria considered in this simulation are both largely penalized by the number of parameters to be

	First simulation			*Second simulation*			
	Euclidean Distance	BIC	AIC	Euclidean Distance	BIC	AIC	Running Times
MN	6.24	8008	7978	7.09	7088	7058	0.001
DM	2.71	5683	5650	6.44	3768	3735	0.005
RCM	3.10	6902	6869	4.52	12151	12118	5.400
NM	5.22	9301	9265	6.13	8251	8215	0.005
GDM	2.58	6075	6015	5.48	4176	4117	0.026
DDM_2	2.49	21430	21328	5.89	16189	160887	1.057
DDM_3	2.26	21436	21297	5.42	16203	16064	2.001
DDM_4	2.14	21437	21263	4.95	16206	16031	2.791
DDM_20	1.69	22032	21277	3.16	16763	16008	23.280

Table 2: Average euclidean distances between the empirical and the estimated variances, BIC, and AIC in the two simulations

estimated. For instance a model with $K = 3$ components involves 83 parameters with respect to the 20 parameters of a simple multinomial distribution. As a consequence, the DDM is never suggested by the two information criteria.

The results of the comparison are shown also from a graphical point of view in Figure 3 that represents the evolution and the trajectory of the empirical true variance - solid black line - with respect to the estimated variances of the different models when the number of zeros in the dataset increases. In both scenarios it is evident the Deep Dirichlet-Multinomial distribution gets a very good approximation that improves as K increases.

In the first simulation study, where the zeros are increasingly inserted in the data in a random way, the empirical variance has a marked parabolic profile. This is due to the fact that the empirical variance increases with the addition of zeros in the data until we get a situation in which half of the data are zeros in the sixth scenario. Here, it is reached the maximum overdispersion and heterogeneity. Then from the seventh scenario, the extra-variation of the data starts to decrease towards the original level because, as the zeros keep increasing, the dataset will tend to be more homogeneous. This reasonable behavior is

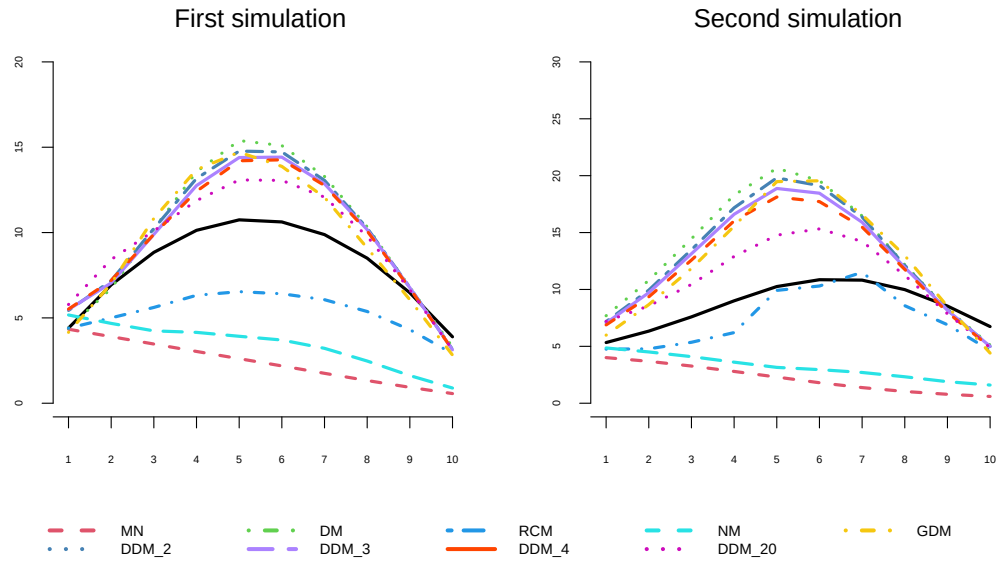

Figure 3: Results of the two simulations. The continuous dark line represents the true empirical variance.

reproduced by the DM, the RCM, the GDM and the DDM but it is the latter that is the closest to the real variance when a consistent overdispersion is present.

A parabolic shape is present also in the second simulation, in which the trajectory of the curves is somehow different due to the different method used in order to add the zeros into the dataset. This, however, does not change the fact that the DDM is a very good model able to describe the empirical variance trajectory under the different scenarios and the approxitation improves as K increases. This aspect will be further discussed in the next section.

4.2. DDM Asymptotic behavior

Exploiting the same data generating process defined for the two simulations above, we analyze the dynamic behavior of the Deep Dirichlet-Multinomial variance as function of K, for the intermediate scenario with a 50% percentage of added zeros.

For each value of $K = 1, \ldots, 50$ the DDM model together with its variance-covariance matrix are estimated with 10 replications each. The summarized results are displayed in Figure 4. The dashed black line represents the sample variance mean across 10 random datasets, while the solid red one describes the evolution of the fitted DDM variance when K increases to $+\infty$. It is straightforward to see how the proposed model is able to account for the extra-variation of the data. In particular, the empirical analysis suggests that the estimated variance tends to the computed variance when the number of elements K of the mixture goes to $+\infty$.

4.3. Choosing K in DDM distribution

In this section we aim to verify whether the optimal number of mixture components of the DDM distribution can be reasonably suggested by the asymptotic criteria AIC and BIC and by the data-driven slope heuristic criterion [33, 34] in which the likelihood is penalized according to the result of the fit of the likelihoods on the complexities of the models.

We considered the two simulation studies with an intermediate proportion of

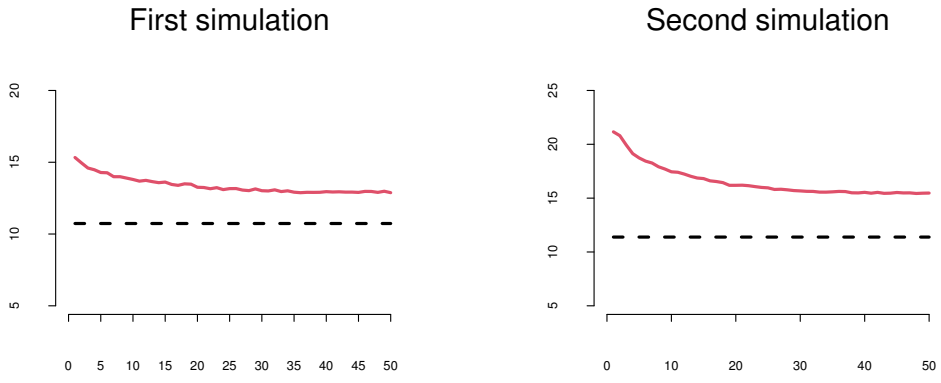

Figure 4: Estimated variability (solid line) vs empirical variability (dashed line) for an increasing number of components.

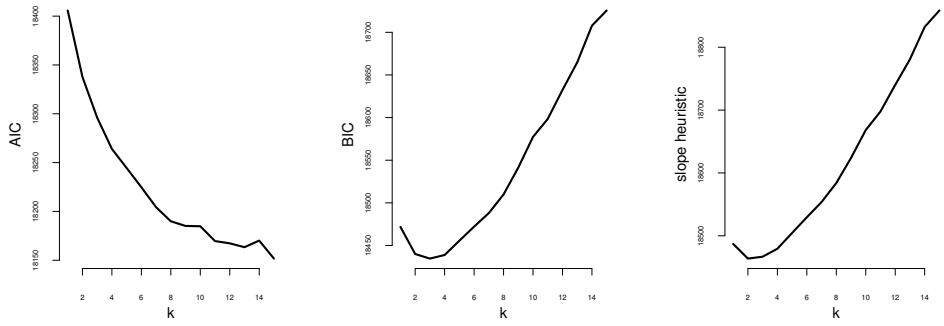

Figure 5: AIC, BIC and slope heuristic (multiplied by -1) for each k in the first simulation

zeros. The DDM model was estimated 10 times for each value $K = 1, \ldots, 15$ of mixture components. Figure 5 and 6 show the AIC, the BIC and the heuristic slopes (multiplied by -1) for each of the two methods of adding the zeros in the dataset.

As expectable in both studies, the BIC is more penalized by the number of parameters with respect to the AIC, which favors many components. According to the BIC, the suggested values of K are 3 and 2 in the two studies; the AIC instead suggests 15 and 12 components in the first and second settings, respectively. The slope heuristic values are consistent with the BIC curve across the different values of K, therefore $K = 2$ or 3 is suggested.

5. An empirical illustration to RNA Sequencing data

In this section an empirical analysis is illustrated on RNA sequencing dataset. Data taken by [35] consist of $p = 714$ microRNAs about cervical cancer that quantify the expression of microRNAs in tumor and non-tumor human cervical tissue samples (see [36] for more details). In this analysis the tumor and non-tumor samples are considered altogether for a total of 58 tissue samples. This leads to an additional source of variability coming from the heterogeneity of the two classes.

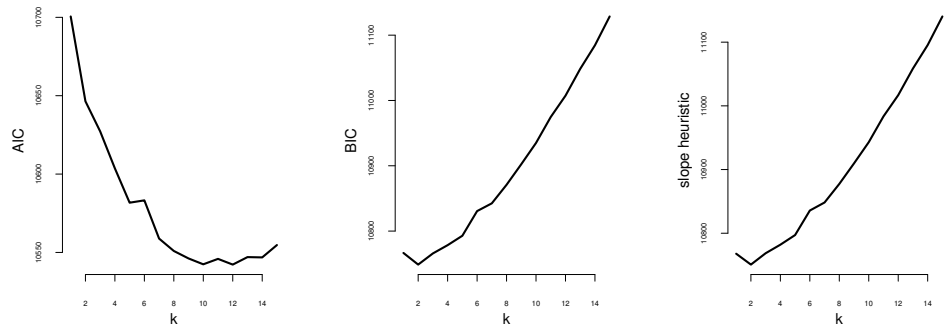

Figure 6: AIC, BIC and slope heuristic (multiplied by -1) for each k in the second simulation

Model	Euclidean Distance
MN	619.562
DM	507.510
RCM	451.199
DDM_2	409.702
DDM_3	409.703
DDM_4	409.704
DDM_20	409.704

Table 3: Euclidean distances between the empirical and the estimated variances of each model.

The Multinomial model (MN), the Dirichlet-Multinomial model (DM), the Random-Clumped model (RCM) and the Deep Dirichlet-Multinomial model (DDM) with different K are estimated on these data and compared in terms of the capability to describe the overall empirical variability. The Negative Multinomial and the Generalized DM cannot be estimated on these data due to their high-dimensionality. Table 3 shows the Euclidean distance between the empirical variances and the estimated variances from each model.

These results confirm the superiority of the DDM model in estimating the computed variances with overdispersion.

6. Final remarks

In this work, we have conducted a comprehensive analysis of the likelihood-based models able to deal with data that present extra-multinomial variation. In addition, a new approach, the Deep Dirichlet-Multinomial distribution, that resembles the deep learning architecture composed by an additional hidden layer with several nodes is proposed. The proposed DMM distribution is characterized by some interesting and desirable properties, although it is not always considered to be the best choice by BIC and AIC due to the large amount of parameters that need to be estimated, compared to the other models analyzed.

First of all, the analytical formula of its variance can be split in two components that ideally represent the within and between variability. This allows to capture both under- and over-dispersion and to have a more flexible correlation structure among variables. Moreover, we showed computationally that the variance of the DDM model tends to the empirical variance when the number of mixture components increases, and this is of course a desirable property that a good distribution should have. The choice of estimating the DDM distribution using an EM algorithm leads to good results in all the simulations considered, even if the price of its large flexibility is in its higher computational times.

References

[1] P. Bach, H. Farbmacher, M. Spindler, Semiparametric count data modeling with an application to health service demand, Econometrics and Statistics 8 (2018) 125–140. doi:https://doi.org/10.1016/j.ecosta.2017.08.004.
URL https://www.sciencedirect.com/science/article/pii/S2452306217300710

[2] K. Poortema, On modelling overdispersion of counts, Stat. Neerl. 53 (1) (1999) 5–20.

[3] B. Efron, Double exponential families and their use in generalized linear regression, Journal of the American Statistical Association 81 (395) (1986) 709–721.

[4] S. Munzert, C. Rubba, P. Meißner, D. Nyhuis, Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining, Wiley, Hoboken, NJ, USA, 2015.

[5] Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics, Nat. Rev. Genet. 10 (2009) 57–63.

[6] D. Lambert, Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing, Technometrics 34 (1) (1992) 1–14.

[7] D. B. Hall, Zero-inflated Poisson and binomial regression with random effects: a case study, Biometrics 56 (4) (2000) 1030–1039. `arXiv:11129458`.

[8] A. O. Diallo, A. Diop, J.-F. Dupuy, Analysis of multinomial counts with joint zero-inflation, with an application to health economics, J. Stat. Plan. Inference 194 (2018) 85–105.

[9] N. D. Y. III, J. R. Wilson, Comparison of quasi-likelihood models for overdispersion, Australian Journal of Statistics 37 (2) (1995) 217–231.

[10] J. G. Morel, A covariance matrix that accounts for different degrees of extraneous variation in multinomial responses, Communications in Statistics-Simulation and Computation 28 (2) (1999) 403–413.

[11] F. Afroz, M. Parry, D. Fletcher, Estimating overdispersion in sparse multinomial data, Biometrics 76 (3) (2020) 834–842.

[12] J. Alonso-Revenga, N. Martín, L. Pardo, New improved estimators for overdispersion in models with clustered multinomial data and unequal cluster sizes, Statistics and Computing 27 (1) (2017) 193–217.

[13] J. E. Mosimann, On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions, Biometrika 49 (1/2) (1962) 65–82.

[14] J. Schmidhuber, Deep learning in neural networks: An overview, Neural Networks 61 (2015) 85–117.

[15] L. D. Valle, F. Leisen, A new multinomial model and a zero variance estimation, Communications in Statistics—Simulation and Computation 39 (4) (2010) 846–859.

[16] I. Olkin, R. Liu, A bivariate beta distribution, Statistics & Probability Letters 62 (4) (2003) 407–412.

[17] J. G. Morel, N. K. Nagaraj, A finite mixture distribution for modelling multinomial extra variation, Biometrika 80 (2) (1993) 363–371.

[18] T. Banerjee, S. Paul, An extension of Morel-Nagaraj's finite mixture distribution for modelling multinomial clustered data, Biometrika 86 (3) (1999) 723–727.

[19] N. K. Neerchal, J. G. Morel, An improved method for the computation of maximum likeliood estimates for multinomial overdispersion models, Computational Statistics & Data Analysis 49 (1) (2005) 33–43.

[20] A. M. Raim, M. Liu, N. K. Neerchal, J. G. Morel, On the method of approximate fisher scoring for finite mixtures of multinomials, Statistical Methodology 18 (2014) 115–130.

[21] O. Bregu, N. Zamzami, N. Bouguila, Mixture-based clustering for count data using approximated fisher scoring and minorization-maximization approaches, Computational Intelligence 37 (1) (2021) 596–620.

[22] Y. Zhang, H. Zhou, J. Zhou, W. Sun, Regression models for multivariate count data, Journal of Computational and Graphical Statistics 26 (1) (2017) 1–13.

[23] Y. Zhang, H. Zhou, MGLM: Multivariate Response Generalized Linear Models, r package version 0.2.0 (2018).
URL https://CRAN.R-project.org/package=MGLM

[24] R. J. Connor, J. E. Mosimann, Concepts of independence for proportions with a generalization of the dirichlet distribution, Journal of the American Statistical Association 64 (325) (1969) 194–206.

[25] D. S. Morris, A. M. Raim, K. F. Sellers, A Conway-Maxwell-multinomial distribution for flexible modeling of clustered categorical data, Journal of Multivariate Analysis 179 (2020) 104651.

[26] P. M. Altham, R. K. Hankin, et al., Multivariate generalizations of the multiplicative binomial distribution: Introducing the MM package, Journal of Statistical Software 46 (12) (2012) 1–23.

[27] C. Viroli, L. Anderlucci, Deep mixtures of unigrams for uncovering topics in textual data, Statistics and Computing 31 (3) (2021) 1–10.

[28] H. D. Nguyen, T. Nguyen, F. Chamroukhi, G. J. McLachlan, Approximations of conditional probability density functions in Lebesgue spaces via mixture of experts models, Journal of Statistical Distributions and Applications 8 (1) (2021) 1–15.

[29] T. T. Nguyen, H. D. Nguyen, F. Chamroukhi, G. J. McLachlan, Approximation by finite mixtures of continuous density functions that vanish at infinity, Cogent Mathematics & Statistics 7 (1) (2020) 1750861. arXiv:https://doi.org/10.1080/25742558.2020.1750861, doi:10.1080/25742558.2020.1750861.
URL https://doi.org/10.1080/25742558.2020.1750861

[30] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society. Series B (Methodological) 39 (1) (1977) 1–38.

[31] G. Schwarz, et al., Estimating the dimension of a model, Annals of statistics 6 (2) (1978) 461–464.

[32] H. Akaike, A new look at the statistical model identification, IEEE transactions on automatic control 19 (6) (1974) 716–723.

[33] L. Birgé, P. Massart, Minimal Penalties for Gaussian Model Selection, Probab. Theory Related Fields 138 (1) (2007) 33–73.

[34] C. Maugis, B. Michel, Data-driven penalty calibration: A case study for Gaussian mixture model selection, ESAIM: Probability and Statistics 15 (2011) 320–339.

[35] D. Witten, R. Tibshirani, S. G. Gu, A. Fire, W.-O. Lui, Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls, BMC biology 8 (1) (2010) 1–14.

[36] D. M. Witten, Classification and clustering of sequencing data using a poisson model, The Annals of Applied Statistics 5 (4) (2011) 2493–2518.