Received: 9 November 2020 | Revised: 24 August 2021 | Accepted: 31 August 2021

DOI: 10.1002/bimj.202000334

Biometrical Journal

RESEARCH ARTICLE

Dirichlet composition distribution for compositional data with zero components: An application to fluorescence in situ hybridization (FISH) detection of chromosome

Man-Lai Tang[1] | **Qin Wu**[2] | **Sheng Yang**[3] | **Guo-Liang Tian**[4]

[1] Department of Mathematics, College of Engineering, Design & Physical Sciences, Brunel University London, Uxbridge, United Kingdom

[2] Department of Statistics, School of Mathematical Sciences, South China Normal University, Guangzhou City, Guangdong, P. R. China

[3] Zhongshan People's Hospital, Zhongshan, P. R. China

[4] Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen City, Guangdong, P. R. China

Correspondence
Qin Wu, Department of Statistics, School of Mathematical Sciences, South China Normal University, Guangzhou City, Guangdong 510631, P. R. China.
Email: wuqin_1985@163.com

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 12171167, 11801184; Research Grant Council of the Hong Kong Special Administrative Region, Grant/Award Numbers: UGC/FDS14/P06/17, UGC/FDS14/P02/18

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to their computational complexity.

Abstract
Zeros in compositional data are very common and can be classified into rounded and essential zeros. The rounded zero refers to a small proportion or below detection limit value, while the essential zero refers to the complete absence of the component in the composition. In this article, we propose a new framework for analyzing compositional data with zero entries by introducing a stochastic representation. In particular, a new distribution, namely the Dirichlet composition distribution, is developed to accommodate the possible essential-zero feature in compositional data. We derive its distributional properties (e.g., its moments). The calculation of maximum likelihood estimates via the Expectation-Maximization (EM) algorithm will be proposed. The regression model based on the new Dirichlet composition distribution will be considered. Simulation studies are conducted to evaluate the performance of the proposed methodologies. Finally, our method is employed to analyze a dataset of fluorescence in situ hybridization (FISH) for chromosome detection.

KEYWORDS
compositional data, Dirichlet distribution, EM algorithm, essential zero, gamma distribution, rounded zeros, stochastic representation

1 | INTRODUCTION

Compositional data, which consist of vectors of positive components subject to a constant-sum constraint (e.g., equal to 1 for proportions and 100 for percentages), record the information about the relative frequencies associated with different components of a system (Ferrers, 1886). They commonly arise in many disciplines such as the components of rocks in geology, the budget share patterns of household expenditures in economics, and the proportion of normal cells in medical research. It is noteworthy that compositional data are subject to the following two intrinsic constraints:

(a) (Bounded support constraint) Each element in the component vector must lie between 0 and 1, inclusive; and
(b) (Summation constraint) All the elements in the component vector must sum to 1.

Mosimann (1962) proposed the Dirichlet-multinomial (DM) distribution, which is a family of discrete multivariate probability distributions on a finite support of nonnegative integers. It is noteworthy that DM is a compound probability distribution, which models counts from a multinomial distribution with a probability vector that is drawn from a Dirichlet distribution. As a result, the DM model (i.e., zero-inflated generalized DM [ZIGDM] by Tang & Chen, 2019) is designed for count data and fails to deal with the compositional data, which are proportions or percentages, described in this article. Applications of statistical methods designed for unconstrained data to such compositional data may result in invalid inference conclusions. For instance, Pearson (1897) discussed the spurious correlation issue in compositional data analysis and concluded that the unit-sum constraint is often intentionally ignored and the statistical methods without constraints are misused, which may eventually lead to disastrous results. Aitchison (1982) first proposed statistical methodology for the compositional data. Aitchison (1982, 1986) first introduced the logistic normal (LN) distribution as a framework for compositional data analyses. In particular, his technique assumes multivariate normality of additive log-ratio transformed data. Since then, various researchers have extended Aitchison's approach in both theoretical and practical respects. For example, Zhang (2000) discussed various distributions for compositional data on the simplex district (e.g., the generalized Dirichlet, additive logistic, and spherical distributions). Egozue et al. (2003) introduced the isometric log-ratio transformation.

In compositional data analysis, the presence of zero components may induce obstacles to the applications of the aforementioned distributional approaches (e.g., zero cannot be the denominator when applying the additive logistic transformation). Aitchison (1986) classified the zeros in compositional data into rounded (or trace elements) zeros and essential (or true) zeros. It is not uncommon that compositional data contain zero components due to either complete absence (i.e., essential zeros), or a small proportion or below the detection limit (i.e., rounded zeros) of certain component(s). Aitchison (1982) pointed out that the log-ratio transformation failed to work when these zeros are denominators.

To deal with the rounded zeros, the most popular method is to replace the rounded zero(s) by a small value (i.e., zero replacement). For example, Palarea-Albaladejo and Martín-Fernández (2008) proposed a modified EM algorithm to replace the rounded zeros in compositional data, Hijazi (2011) developed the EM algorithm–based method to deal with rounded zeros. The nonparametric imputation approach is proposed by Martín-Fernández et al. (2003) to handle rounded zeros.

For essential zeros, three well-known approaches have been developed. The data amalgamation, which was proposed by Aitchison (1990), is to eliminate the components with zero elements by combining them with some other nonzero components. The second approach models the zeros separately (e.g., Aitchison, 1986; Bear & Billheimer, 2016; Zadora et al., 2010). For instance, Bear and Billheimer (2016) projected compositions with zeros onto smaller dimensional subspaces. As a result, they developed a mixture of logistic normals which successfully addresses the issues of division by zero and the log of zero. The third approach is to transform compositions into directional data on the hypersphere and develop a regression model using the Kent distribution (e.g., Kent, 1982; Scealy & Welsh, 2011), which tolerates zeros. Other methods are also investigated, such as the mixture models to eliminate the essential zeros (Stewart & Field, 2011), the latent Gaussian model (Butler & Glasbey, 2008), and the Dirichlet regression model (Tsagris & Stewart, 2018).

In clinical research, compositional data with essential zeros are not uncommon. For instance, chromosome abnormalities are considered to be the most common cause of spontaneous abortion. Fluorescence in situ hybridization (FISH) is a cytogenetic technique developed in the early 1980s (see, e.g., Langer-Safer et al., 1982). It uses fluorescent DNA probes to target specific chromosomal locations within the nucleus, resulting in colored signals that can be detected using a fluorescent microscope. For spontaneous abortion, a damaged embryo is taken out from the gravida and the FISH technique is then employed to detect the cells which are selected randomly from the damaged embryo. Finally, the respective proportions of diploidy, triploidy, and polyploidy at chromosome 22 for those randomly chosen and tested cells are recorded for each embryo. Obviously, the observations are compositional data (i.e., total sum is equal to one). For example, an observation of (0.2,0.3,0.5) means 20%, 30%, and 30% of the selected cells are chromosome diploidy, chromosome triploidy, and chromosome polyploidy, respectively. The FISH data reported in the Supporting Information are the compositional observations of 51 embryos from the curettage operation in Zhongshan People's Hospital in Mainland China. The age of each gravida is also reported. It is noteworthy that nearly 80% (i.e., 40 out of 51) of the embryos demonstrate purely normal chromosomes (i.e., compositional observation being $(1, 0, 0)$). Most importantly, none of the aforementioned approaches are suitable for our FISH data, which motivate the present article.

The rest of this paper is organized as follows. In Section 2, we introduce a new stochastic representation (SR) for compositional data with zero components and the new Dirichlet composition distribution (DCD) is defined. Likelihood-based methods for parameter estimation and confidence intervals construction without covariates will be provided in Section 3. Regression model analysis based on the distribution will be considered in Section 4. Simulation studies will be conducted to examine the performance of our proposed methods in Section 5. We will revisit and analyze the FISH dataset in Section 6. A brief discussion will be presented in Section 7. Some technical details are included in the Appendix.

2 | NEW DEFINITION OF COMPOSITIONAL RANDOM VECTOR AND THE DCD

In this section, we introduce a new definition of a compositional random vector which can be adopted for modeling the compositional data. The definition is proposed based on SR. We then introduce the DCD by assuming the base vector following independent Gamma distributions.

2.1 | Definition of a compositional random vector

To model the zero elements in the compositional data, we employ the SR to establish the definition of a compositional random vector.

Definition 1. (Compositional random vector). A random vector $\mathbf{X} = (X_1, \dots, X_m)^\top$ is said to be an m-dimensional compositional random vector if

$$\mathbf{X} \overset{\mathrm{d}}{=} \frac{\mathbf{Z} \circ \mathbf{Y}}{\mathbf{Z}^\top \mathbf{Y}} \tag{1}$$

with $\mathbf{Z} = (Z_1, \dots, Z_m)^\top$ being the indicator vector (i.e., $Z_j = 0$ or 1 for $j = 1, \dots, m$) such that $\sum_{j=1}^m Z_j \neq 0$, and $\mathbf{y} = (Y_1, \dots, Y_m)^\top$ being the base vector with each element being positive random variable (i.e., $Y_j \in (0, +\infty)$, for $j = 1, \dots, m$), "$\overset{\mathrm{d}}{=}$" meaning the random variables on both sides have the same distribution, $\mathbf{Z} \circ \mathbf{Y} = (Z_1 Y_1, \dots, Z_m Y_m)^\top$ and $\mathbf{Z}^\top \mathbf{Y} = \sum_{i=1}^n Z_i Y_i$.

The indicator vector $\mathbf{Z}$ provides the possibility of zero entries in the distribution with $Z_j = 0$ meaning the jth component in $\mathbf{X}$ being zero. The base vector $\mathbf{Y}$ carries the quantitative information. It can be any positive vector and determines the nonzero components in $\mathbf{X}$.

2.2 | Definition of DCD

In the compositional random vector, if we let $\mathbf{Z}$ be the m-dimensional independent Bernoulli random variables by excluding the point $\mathbf{0}$, and $\mathbf{Y}$ be the independent Gamma random variable with different shape parameters α_i but identical rate parameter β, we can define a new distribution called DCD. Since the rate parameter β will be eliminated in the SR of the compositional random vector, β is unidentifiable in the distribution. Without loss of generality, we assume $\beta = 1$. That is, for each $\mathbf{Y} = (Y_1, \dots, Y_m)^\top$, we have $\{Y_i\}_{i=1}^m \overset{\mathrm{ind}}{\sim} \mathrm{Gamma}(\alpha_i, 1)$.

Definition 2. (DCD). A compositional random vector $\mathbf{X} \overset{\mathrm{d}}{=} \frac{\mathbf{Z} \circ \mathbf{Y}}{\mathbf{Z}^\top \mathbf{Y}}$ is said to follow the DCD, denoted by $\mathrm{DCD}(\boldsymbol{p}, \boldsymbol{\alpha})$, if $\mathbf{Z} \perp\!\!\!\perp \mathbf{Y}$ and

$$\Pr(\mathbf{Z} = \boldsymbol{z}) = \frac{\prod_{j=1}^m (1 - p_j)^{z_j} p_j^{1 - z_j}}{1 - p_1 \cdots p_m}, \tag{2}$$

and

$$f(\mathbf{Y} = \mathbf{y}) = \prod_{i=1}^{m} \frac{\mathrm{e}^{y_i} y_i^{\alpha_i - 1}}{\Gamma(\alpha_i)}, \tag{3}$$

where $\boldsymbol{p} = (p_1, \dots, p_m)^\top$ contains the parameters of the indicator vector and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m)^\top$ contains the parameters of the base vector with $\{0 \le p_j < 1, \alpha_j > 0\}_{j=1}^{m}$.

The probability density function of $\mathbf{X}$ is then given by

$$f(\boldsymbol{x}) = \frac{\prod_{j=1}^{m} p_j^{1-z_j}(1-p_j)^{z_j}}{1 - p_1 \cdots p_m} \times \left[\frac{\Gamma(\alpha_{\mathbb{J}}^*)}{\prod_{j \in \mathbb{J}} \Gamma(\alpha_j)} \prod_{j \in \mathbb{J}} x_j^{\alpha_j - 1} \right], \tag{4}$$

where $z_j = I(x_j > 0), j = 1, \dots, m$ and $\alpha_{\mathbb{J}}^* = \sum_{j \in \mathbb{J}} \alpha_j$, $\mathbb{J}$ is the subset of the index with $\boldsymbol{x}$ being positive (i.e., $x_j > 0$ for any $j \in \mathbb{J}$ and $x_j = 0$ for $j \notin \mathbb{J}$). (For more details of the probability density function, refer to Appendix A.1.)

Remark 1. It is clear that $p_j \neq 1$ for $j = 1, \dots, m$. As $\Pr(Z_j = 0) = p_j = 1$ implies that the element in the jth column must be 0, we can simply delete the jth column and the remaining $m - 1$ columns still form a compositional random vector.

Remark 2. If $p_j = 0$ for $j = 1, \dots, m$, it means $Z_j = 1$, we have DCD($\mathbf{0}, \boldsymbol{\alpha}$)=Dirichlet($\alpha_1, \dots, \alpha_m$). That is, the well-known Dirichlet distribution is a special case of DCD($\boldsymbol{p}, \boldsymbol{\alpha}$).

Remark 3. We here suppose $\mathbf{Z}$ follows the zero-truncated multivariate Bernoulli distribution. Due to SR in (1), the denominator $\mathbf{Z}^\top \mathbf{Y} = \sum_{i=1}^{n} Z_i Y_i$ must be nonzero; therefore, $Z_1, \dots, Z_m$ are not independent as they cannot be 0 at the same time. That is, $\{Z_j\}_{j=1}^{m}$ follow the independent Bernoulli distributions but exclude the point $\mathbf{0}$.

2.3 | Mixed moments and moment generating function

If $\mathbf{X} \sim$ DCD($\boldsymbol{p}, \boldsymbol{\alpha}$), then the following results can be easily shown:

$$E(\mathbf{X}) = \sum_{\boldsymbol{z}} \frac{\prod_{j=1}^{m} p_j^{1-z_j}(1-p_j)^{z_j}}{1 - p_1 \cdots p_m} \times \left(\mathbf{0}_{\mathbb{K}}, \frac{\boldsymbol{\alpha}_{\mathbb{J}}}{\alpha_{\mathbb{J}}^*} \right)^\top,$$

$$\mathrm{Cov}(\mathbf{X}, \mathbf{X}) = \sum_{\boldsymbol{z}} \frac{\prod_{j=1}^{m} p_j^{1-z_j}(1-p_j)^{z_j}}{1 - p_1 \cdots p_m} \times \mathrm{Cov}(\mathbf{X}|\mathbf{Z}, \mathbf{X}|\mathbf{Z}), \tag{5}$$

$$E\Big(\prod_{j \in \mathbb{J}} X_j^{r_j}\Big) = \frac{B(\sum_{j \in \mathbb{J}} \alpha_j + \sum_{j \in \mathbb{J}} r_j)}{B(\sum_{j \in \mathbb{J}} \alpha_j)}; \quad E\Big(\prod_{\exists j \notin \mathbb{J}} X_j^{r_j}\Big) = 0, \tag{6}$$

where $\mathbb{K}$ is the subset with $\boldsymbol{x}$ being 0 (i.e., $\mathbb{K} = \{1, \dots, m\} \backslash \mathbb{J}$), $(\mathbf{0}_{\mathbb{K}}, \frac{\alpha_{\mathbb{J}}}{\alpha_{\mathbb{J}}^*})^\top = (e_i)_{m \times 1}$, and $\mathrm{Cov}(\mathbf{X}, \mathbf{X}) = (v_{ij})_{m \times m}$

$$\begin{array}{rcll} e_i = & 0; & \text{if} & i \in \mathbb{K}; \\ e_i = & \frac{\alpha_i}{\alpha_{\mathbb{J}}^*}; & \text{if} & i \in \mathbb{J}; \\ v_{ii} = & 0; & \text{if} & i \in \mathbb{K}; \end{array}$$

$$\begin{aligned}
v_{ii} &= \frac{\alpha_i(\alpha^*_{\mathbb{J}}-\alpha_i)}{(\alpha^*_{\mathbb{J}})^2(\alpha^*_{\mathbb{J}}+1)}, \quad \text{if} \quad i \in \mathbb{J};\\
v_{ij} &= \quad 0; \quad \text{if} \quad i \in \mathbb{K} \quad \text{or} \quad j \in \mathbb{K};\\
v_{ij} &= -\frac{\alpha_i\alpha_j}{(\alpha^*_{\mathbb{J}})^2(\alpha^*_{\mathbb{J}}+1)}, \quad \text{if} \quad i, j \in \mathbb{J}.
\end{aligned} \tag{7}$$

3 | Statistical inference without covariates

In this section, we present statistical inferences based on data without covariates, which include the maximum likelihood estimates (MLEs) calculation in Section 3.1 and the confidence interval construction in Section 3.2.

3.1 | Maximum likelihood estimates for target parameters

Suppose the n observations are $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$, where $\boldsymbol{x}_i = (x_{i1}, \dots, x_{im})^\top$ and m is the number of dimensions. Without loss of information, we assume that there are zero entries in the first k observations, that is $\min(\boldsymbol{x}_i) = 0$ for $i = 1, \dots, k$, and $\min(\boldsymbol{x}_i) > 0$ for $i = k+1, \dots, n$, $0 \le k \le m$. We have $\mathbb{J}_i = (1, 2, \dots, m)^\top$ when $i \ge k$. Therefore, the observed likelihood function is given by

$$\begin{aligned}
L(\boldsymbol{p}, \boldsymbol{\alpha} | Y_{\text{obs}}) = &\left\{ \prod_{i=1}^{n} \frac{\prod_{j=1}^{m} p_j^{1-z_{ij}}(1-p_j)^{z_{ij}}}{1 - p_1 \cdots p_m} \right\} \times \left\{ \prod_{i=1}^{k} \left[\frac{\Gamma(\alpha^*_{\mathbb{J}_i})}{\prod_{j\in\mathbb{J}_i} \Gamma(\alpha_j)} \prod_{j\in\mathbb{J}_i} x_{ij}^{\alpha_j - 1} \right] \right\}\\
&\times \left\{ \prod_{i=k+1}^{n} \left[\frac{\Gamma(\sum_{j=1}^{m} \alpha_j)}{\prod_{j=1}^{m} \Gamma(\alpha_j)} \prod_{j=1}^{m} x_{ij}^{\alpha_j - 1} \right] \right\},
\end{aligned} \tag{8}$$

where $\mathbb{J}_i$ denotes the index set of those positive elements in each $\boldsymbol{x}_i$ (i.e., $x_{ij} > 0$ if $j \in \mathbb{J}_i$). Here, the indicator variable $\mathbf{Z}$ can be observed via the observation $\boldsymbol{x}_i$ as

$$Z_{ij} | (\mathbf{X}_i = \boldsymbol{x}_i) = \begin{cases} 0, & \text{if} \quad x_{ij} = 0,\\ 1, & \text{if} \quad x_{ij} > 0. \end{cases} \tag{9}$$

Observing that $Z_j = 1$ is equivalent to $j \in \mathbb{J}_i$. we have

$$\begin{aligned}
l(\boldsymbol{p}, \boldsymbol{\alpha} | Y_{\text{obs}}) = &\left\{ \sum_{i=1}^{n} \left[(1 - z_{ij}) \log p_j + z_{ij} \log(1 - p_j) - \log(1 - p_1 \cdots p_m) \right] \right\}\\
&+ \sum_{i=1}^{k} \left\{ \log \Gamma(\alpha^*_{\mathbb{J}_i}) + \sum_{j\in\mathbb{J}_i} \left[(\alpha_j - 1) \log x_{ij} - \log \Gamma(\alpha_j) \right] \right\}\\
&+ \sum_{i=k+1}^{n} \left\{ \log \Gamma(\sum_{j=1}^{m} \alpha_j) + \sum_{j=1}^{m} \left[(\alpha_j - 1) \log x_{ij} - \log \Gamma(\alpha_j) \right] \right\}.
\end{aligned} \tag{10}$$

Instead of obtaining the MLEs of the parameters $\boldsymbol{p} = (p_1, \dots, p_m)^\top$ and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m)^\top$ via solving the solutions to the system of equations $\frac{\partial l(\boldsymbol{p}, \boldsymbol{\alpha} | Y_{\text{obs}})}{\partial \boldsymbol{p}} = \mathbf{0}$ and $\frac{\partial l(\boldsymbol{p}, \boldsymbol{\alpha} | Y_{\text{obs}})}{\partial \boldsymbol{\alpha}} = \mathbf{0}$, we consider the EM algorithm. Motivated by the SR, we introduce the base vectors $\{\mathbf{y}_i\}_{i=1}^{n}$ and s as missing data, where s denotes the number of unobserved $\mathbf{0}$ to make the components in $\mathbf{Z}$ being independent. In fact, $\mathbf{Z}$ are independent Bernoulli variables which exclude the outcome of $\mathbf{0}$. Therefore,

$\{\underbrace{\mathbf{0},\dots,\mathbf{0}}_{s},\{\boldsymbol{z}_i\}_{i=1}^{n}\}$ are the complete observations and the likelihood function based on the complete data is given by

$$L_{com}(\boldsymbol{p},\boldsymbol{\alpha}|Y_{\text{com}}) = (p_1\cdots p_m)^s \times \left\{\prod_{i=1}^{n}\prod_{j=1}^{m}\left[p_j^{1-z_{ij}}(1-p_j)^{z_{ij}} \times \frac{\mathrm{e}^{y_{ij}}y_{ij}^{\alpha_j-1}}{\Gamma(\alpha_j)}\right]\right\}, \tag{11}$$

and the log-likelihood function of the complete data likelihood function is

$$\begin{aligned} l_{com}(\boldsymbol{p},\boldsymbol{\alpha}|Y_{\text{com}}) &= s\sum_{j=1}^{m}\log p_j + \sum_{i=1}^{n}\sum_{j=1}^{m}\left[z_{ij}\log(1-p_j)+(1-z_{ij})\log p_j\right] \\ &\quad + \left\{\sum_{i=1}^{n}\sum_{j=1}^{m}\left[-y_{ij}+(\alpha_j-1)\log y_{ij}-\log\Gamma(\alpha_j)\right]\right\}. \end{aligned} \tag{12}$$

The M-step is to solve the following equations, for $j = 1,\dots,m$:

$$\begin{cases} \dfrac{\partial l_{com}(\boldsymbol{p},\boldsymbol{\alpha}|Y_{\text{com}})}{\partial p_j} &= \dfrac{s}{p_j} + \displaystyle\sum_{i=1}^{n}\left(\frac{1-z_{ij}}{p_j} - \frac{z_{ij}}{1-p_j}\right) = 0, \\ \dfrac{\partial l_{com}(\boldsymbol{p},\boldsymbol{\alpha}|Y_{\text{com}})}{\partial \alpha_j} &= \displaystyle\sum_{i=1}^{n}[\log y_{ij} - \psi(\alpha_j)] = 0, \end{cases} \tag{13}$$

where ψ denotes the digamma function with $\psi(\alpha) = \frac{d\log(\Gamma(\alpha))}{d\alpha} = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$. For $p_j, j = 1,\dots,m$, we have

$$p_j = 1 - \frac{\sum_{i=1}^{n} z_{ij}}{n+s}. \tag{14}$$

However, there are no closed-form solutions for α_js and we will use the following Newton–Raphson iterative algorithm to find the MLEs of $\boldsymbol{\alpha}$s.

$$\alpha_j^{(t+1)} = \alpha_j^{(t)} - \left[\frac{\partial l_{com}^2(\boldsymbol{p},\boldsymbol{\alpha}|Y_{\text{com}})}{\partial \alpha_j^2}\right]^{-1} \times \frac{\partial l_{com}(\boldsymbol{p},\boldsymbol{\alpha}|Y_{\text{com}})}{\partial \alpha_j}, \tag{15}$$

where $\frac{\partial l_{com}^2(\boldsymbol{p},\boldsymbol{\alpha}|Y_{\text{com}})}{\partial \alpha_j^2} = -n\phi(\alpha_j)$ with $\phi(\alpha_j) = \frac{d^2\log(\Gamma(\alpha))}{d\alpha^2} = \frac{d\psi(\alpha)}{d\alpha}$ being the trigamma function.

To obtain the E step, we have the following theorem and the proof is presented in Appendix A.2.

Theorem 1. *The conditional expectation of* $\log y_{ij}$ *given* $\boldsymbol{x}$ *is as follows:*

$$E(\log y_j|\boldsymbol{x}) = \begin{cases} \psi(\alpha_j), & \textit{if} \quad x_j = 0, \\ \psi(\alpha_{\mathbb{J}_i}^*) + \log x_j & \textit{if} \quad x_j > 0. \end{cases} \tag{16}$$

The E-step is to replace the missing data by the following conditional expectations:

$$\begin{aligned} E(s) &= \frac{np_1\cdots p_m}{1-p_1\cdots p_m} \\ E(\log y_j|\boldsymbol{x}) &= \begin{cases} \psi(\alpha_j), & \text{if} \quad x_j = 0, \\ \psi(\alpha_{\mathbb{J}_i}^*) + \log x_j & \text{if} \quad x_j > 0. \end{cases} \end{aligned} \tag{17}$$

Here, we can consider the initial values of parameters being $\boldsymbol{\alpha}^0 = (1, \dots, 1)^\top$ in the EM algorithm. The above steps (i.e., E- and M-steps) are repeated until a certain convergence condition is achieved. For instance, if the difference between two successive log-likelihood values is less than the prespecified value 0.001, the algorithm stops after 100–150 iterations.

3.2 | Confidence interval construction

In this section, we will consider the construction of confidence intervals for target parameters using the bootstrap method. It is noted that the value of p_j must be restricted within the interval [0,1]. However, Wald-type confidence intervals may produce upper (or lower) limit larger (or less) than 1 (or 0). It is noteworthy that the MLE of $\boldsymbol{p}$ obtained via our proposed EM algorithm always lies between 0 and 1. As a result, we apply the bootstrap method to create the bootstrap confidence interval (CI) for any arbitrary function of $\boldsymbol{\theta} = (\boldsymbol{p}, \boldsymbol{\alpha})$, denoted by $\vartheta = h(\boldsymbol{\theta})$. Briefly, based on the observations, we can independently generate $\{\mathbf{y}_i\}_{i=1}^n$ with each $\{\mathbf{y}_i\}$ is randomly selected from the n observations with replacement. Having obtained $Y_{\text{obs}}^* = \{\boldsymbol{y}_1^*, \dots, \boldsymbol{y}_n^*\}$, we can calculate the parameter estimates $\hat{\boldsymbol{\theta}}^*$ and get the bootstrap replication $\hat{\vartheta}^* = h(\hat{\boldsymbol{\theta}}^*)$. Independently repeating this process B times, we obtain B replications $\{\hat{\vartheta}_g^*\}_{g=1}^B$. The bootstrap CI of ϑ can be constructed by $[\vartheta_{\text{L}}, \vartheta_{\text{U}}]$, where ϑ_{L} and ϑ_{U} are the $100(\alpha/2)$ and $100(1-\alpha/2)$ percentiles of $\{\hat{\vartheta}_g^*\}_{g=1}^B$, respectively.

4 | STATISTICAL INFERENCE WITH COVARIATES

In this section, we will show how to formulate the regression model for the target parameters and how to obtain the MLEs of the coefficients in the regression model. Let the covariates of each observation be denoted by $\{\boldsymbol{v}_i, \boldsymbol{w}_i\}$, $i = 1, \dots, n$. We consider the following regression models:

$$\begin{cases} \mathbf{X}_i & \overset{\text{ind}}{\sim} & \text{DCD}(\boldsymbol{p}_i, \boldsymbol{\alpha}_i), \qquad i = 1, \dots, n, \\ \log\left(\dfrac{p_{ij}}{1 - p_{ij}}\right) & = & \boldsymbol{v}_i^\top \boldsymbol{\beta}_j, \text{ and} \\ \log(\alpha_{ij}) & = & \boldsymbol{w}_i^\top \boldsymbol{\gamma}_j, \qquad j = 1, \dots, m. \end{cases} \tag{18}$$

Let s_i denote the number of supplementary $\mathbf{0}$ to make the elements in $\boldsymbol{z}_i$ being independent, where $i = 1, \dots, n$. Obviously, $s_1, \dots, s_n$ are missing data with $s_i = 1$ being equivalent to $Z_{i1} = \cdots Z_{im} = 0$ and $\Pr(s_i = 1) = p_{i1} \cdots p_{im}$. Thus, the complete likelihood function is given by

$$\begin{aligned} L_1(\boldsymbol{\beta}, \boldsymbol{\gamma} | Y_{\text{com}}) &= \prod_{i=1}^n \prod_{j=1}^m \left[p_{ij}^{1-z_{ij}} (1 - p_{ij})^{z_{ij}} (p_{i1} \cdots p_{im})^{s_i} \times \frac{\mathrm{e}^{-y_{ij}} y_{ij}^{\alpha_{ij}-1}}{\Gamma(\alpha_{ij})} \right] \\ &= \prod_{i=1}^n \prod_{j=1}^m \left[\left(\frac{\mathrm{e}^{\boldsymbol{v}_i^\top \boldsymbol{\beta}_j}}{1 + \mathrm{e}^{\boldsymbol{v}_i^\top \boldsymbol{\beta}_j}} \right)^{1-z_{ij}} \left(\frac{1}{1 + \mathrm{e}^{\boldsymbol{v}_i^\top \boldsymbol{\beta}_j}} \right)^{z_{ij}} \left(\prod_{l=1}^m \frac{\mathrm{e}^{\boldsymbol{w}_i^\top \boldsymbol{\beta}_l}}{1 + \mathrm{e}^{\boldsymbol{w}_i^\top \boldsymbol{\beta}_l}} \right)^{s_i} \times \frac{\mathrm{e}^{-y_{ij}} y_{ij}^{\mathrm{e}^{\boldsymbol{w}_i^\top \boldsymbol{\gamma}_j} - 1}}{\Gamma(\mathrm{e}^{\boldsymbol{w}_i^\top \boldsymbol{\beta}_j})} \right]. \end{aligned} \tag{19}$$

Or, the log-likelihood function is

$$\begin{aligned} l_1(\boldsymbol{\beta}, \boldsymbol{\gamma} | Y_{\text{com}}) = {} & \sum_{i=1}^n \sum_{j=1}^m \left[(1 - z_{ij} + s_i) \log p_{ij} + z_{ij} \log(1 - p_{ij}) \right] \\ & + \sum_{i=1}^n \sum_{j=1}^m \left[-y_{ij} + (\alpha_{ij} - 1) \log(y_{ij}) - \log(\Gamma(\alpha_{ij})) \right] \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n}\sum_{j=1}^{m}\left\{(1+s_i)\left[\boldsymbol{v}_i^\top\boldsymbol{\beta}_j - \log\left(1+\mathrm{e}^{\boldsymbol{v}_i^\top\boldsymbol{\beta}_j}\right)\right] - z_{ij}\boldsymbol{v}_i^\top\boldsymbol{\beta}_j\right\}\\
&\quad + \sum_{i=1}^{n}\sum_{j=1}^{m}\left[-y_{ij} + (\mathrm{e}^{\boldsymbol{w}_i^\top\boldsymbol{\gamma}_j}-1)\log(y_{ij}) - \log(\Gamma(\mathrm{e}^{\boldsymbol{w}_i^\top\boldsymbol{\gamma}_j}))\right]\\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}\left[(1+s_i-z_{ij})\boldsymbol{v}_i^\top\boldsymbol{\beta}_j - (1+s_i)\log\left(1+\mathrm{e}^{\boldsymbol{v}_i^\top\boldsymbol{\beta}_j}\right)\right]\\
&\quad + \sum_{i=1}^{n}\sum_{j=1}^{m}\left[-y_{ij} + (\mathrm{e}^{\boldsymbol{w}_i^\top\boldsymbol{\gamma}_j}-1)\log(y_{ij}) - \log(\Gamma(\mathrm{e}^{\boldsymbol{w}_i^\top\boldsymbol{\gamma}_j}))\right]. \qquad (20)
\end{aligned}
$$

The MLEs of the regression coefficients are the solution to the following equations:

$$
\frac{\partial l_1(\boldsymbol{\beta},\boldsymbol{\gamma}|Y_{com})}{\partial\boldsymbol{\beta}} = \mathbf{0} \text{ and } \frac{\partial l_1(\boldsymbol{\beta},\boldsymbol{\gamma}|Y_{com})}{\partial\boldsymbol{\gamma}} = \mathbf{0}. \qquad (21)
$$

It is obvious that there is no closed-form solution to (21). Here, we use the Newton–Raphson algorithm to calculate the MLEs, and the iterations are given by

$$
\begin{aligned}
\boldsymbol{\beta}_j^{(t+1)} &= \boldsymbol{\beta}_j^{(t)} - \left[\frac{\partial l_1^2(\boldsymbol{\beta},\boldsymbol{\gamma}|Y_{\mathrm{com}})}{\partial\boldsymbol{\beta}_j\boldsymbol{\beta}_j^\top}\right]^{-1} \times \frac{\partial l_1(\boldsymbol{\beta},\boldsymbol{\gamma}|Y_{\mathrm{com}})}{\partial\boldsymbol{\beta}_j} \text{ and}\\
\boldsymbol{\gamma}_j^{(t+1)} &= \boldsymbol{\gamma}_j^{(t)} - \left[\frac{\partial l_1^2(\boldsymbol{\beta},\boldsymbol{\gamma}|Y_{\mathrm{com}})}{\partial\boldsymbol{\gamma}_j\boldsymbol{\gamma}_j^\top}\right]^{-1} \times \frac{\partial l_1(\boldsymbol{\beta},\boldsymbol{\gamma}|Y_{\mathrm{com}})}{\partial\boldsymbol{\gamma}_j}, \qquad j=1,\dots,m. \qquad (22)
\end{aligned}
$$

The first and negative second partial derivatives of the complete-data log-likelihood function are given by

$$
\begin{aligned}
\frac{\partial l_1(\boldsymbol{\beta},\boldsymbol{\gamma}|Y_{\mathrm{com}})}{\partial\boldsymbol{\beta}_j} &= \sum_{i=1}^{n}[(1-z_{ij}-p_{ij})\boldsymbol{v}_i + s_i(1-p_{ij})\boldsymbol{v}_i]\\
&= \mathbf{V}^\top[(\mathbf{1}-\boldsymbol{z}_{(j)}-\boldsymbol{p}_{(j)}) + (\mathbf{1}-\boldsymbol{p}_{(j)})\circ\boldsymbol{s}],\\
\frac{\partial l_1(\boldsymbol{\beta},\boldsymbol{\gamma}|Y_{\mathrm{com}})}{\partial\boldsymbol{\gamma}_j} &= \sum_{i=1}^{n}\left[\alpha_{ij}\log y_{ij} - \psi(\alpha_{ij})\alpha_{ij}\right]\boldsymbol{w}_i = \mathbf{W}^\top[\boldsymbol{\alpha}_{(j)}\circ(\log(\boldsymbol{y}_{(j)}) - \psi(\boldsymbol{\alpha}_{(j)}))],\\
-\frac{\partial^2 l_1(\boldsymbol{\beta},\boldsymbol{\gamma}|Y_{\mathrm{com}})}{\partial\boldsymbol{\beta}_j\partial\boldsymbol{\beta}_j^\top} &= \sum_{i=1}^{n}[p_{ij}(1-p_{ij})(1+s_i)\boldsymbol{v}_i\boldsymbol{v}_i^\top] \qquad (23)\\
&= \mathbf{V}^\top\mathrm{diag}[\boldsymbol{p}_{(j)}\circ(\mathbf{1}-\boldsymbol{p}_{(j)})\circ(\mathbf{1}+\boldsymbol{s}]\mathbf{V}\\
&\hat{=}\ \mathbf{J}_{\mathrm{com}}(\boldsymbol{\beta}_j), \text{ and}\\
-\frac{\partial^2 l_1(\boldsymbol{\beta},\boldsymbol{\gamma}|Y_{\mathrm{com}})}{\partial\boldsymbol{\gamma}_j\partial\boldsymbol{\gamma}_j^\top} &= \sum_{i=1}^{n}\left[\alpha_{ij}\log y_{ij} - \phi(\alpha_{ij})\alpha_{ij}^2 - \psi(\alpha_{ij})\alpha_{ij}\right]\boldsymbol{w}_i\boldsymbol{w}_i^\top\\
&= -\mathbf{W}^\top\mathrm{diag}\left[(\boldsymbol{\alpha}_{(j)}\circ\log(\boldsymbol{y}_{(j)}) - \boldsymbol{\Phi}(\boldsymbol{\alpha}_{(j)})\circ\boldsymbol{\alpha}_{(j)}^2 - \boldsymbol{\Psi}(\boldsymbol{\alpha}_{(j)})\circ\boldsymbol{\alpha}_{(j)})\right]\mathbf{W}\\
&\hat{=}\ \mathbf{J}_{\mathrm{com}}(\boldsymbol{\gamma}_j),
\end{aligned}
$$

where

$$\begin{array}{rclrclrcl} \mathbf{V} &=& (\boldsymbol{v}_1,\dots,\boldsymbol{v}_n)^\top, & \boldsymbol{z}_{(j)} &=& (z_{1j},\dots,z_{nj})^\top, & \boldsymbol{p}_{(j)} &=& (p_{1j},\dots,p_{nj})^\top, \\ \mathbf{W} &=& (\boldsymbol{w}_1,\dots,\boldsymbol{w}_n)^\top & \boldsymbol{s} &=& (s_1,\dots,s_n)^\top, & \boldsymbol{\alpha}_{(j)} &=& (\alpha_{1j},\dots,\alpha_{nj})^\top, \\ \boldsymbol{y}_{(j)} &=& (y_{1j},\dots,y_{nj})^\top, & j &=& 1,\dots,m. \end{array} \tag{24}$$

Here, $\mathbf{J}_{\rm com}(\boldsymbol{\beta}_i)$ and $\mathbf{J}_{\rm com}(\boldsymbol{\gamma}_i)$ are actually the complete-data Fisher information matrices associated with the parameter vectors $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_j$, respectively, which depend on neither the observed data nor the latent/missing data.

To obtain the MLEs of the parameter vectors $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_j$ in the presence of missing data (i.e., $s_1,\dots,s_n$), we introduce the EM algorithm. Briefly, the M-step is to separately calculate the MLEs of $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_j$ via Newton–Raphson algorithms as follows:

$$\begin{cases} \boldsymbol{\beta}_j^{(t+1)} &=& \boldsymbol{\beta}_j^{(t)} + \mathbf{J}_{\rm com}^{-1}(\boldsymbol{\beta}_j^{(t)})\mathbf{V}^\top\Big[(\mathbf{1} - \boldsymbol{z}_{(j)} - \boldsymbol{p}_{(j)}) + (1 - \boldsymbol{p}_{(j)})\boldsymbol{s}\Big], \text{ and} \\ \boldsymbol{\gamma}_j^{(t+1)} &=& \boldsymbol{\gamma}_j^{(t)} + \mathbf{J}_{\rm com}^{-1}(\boldsymbol{\gamma}_i^{(t)})\mathbf{W}^\top\big[\boldsymbol{\alpha}_{(j)} \circ \log(\boldsymbol{y}_{(j)}) - \psi(\boldsymbol{\alpha}_{(j)})\big], \qquad j = 1,\dots,m. \end{cases} \tag{25}$$

The E-step is to replace s_i in (25) by their conditional expectations, that is,

$$E(s_i|Y_{\rm obs},\boldsymbol{\beta},\boldsymbol{\gamma}) = \left(\prod_{j=1}^m p_{ij}\right)\Bigg/\left(1 - \prod_{j=1}^m p_{ij}\right), \tag{26}$$

$$E(\log y_{ij}|Y_{\rm obs},\boldsymbol{\beta},\boldsymbol{\gamma}) \overset{(6)}{=} \begin{cases} \psi(\alpha_{ij}), & \text{if} \quad x_{ij} = 0, \\ \psi(\sum\limits_{j\in J_i}\alpha_{ij}) + \log x_{ij}, & \text{if} \quad x_{ij} > 0 \end{cases}, \tag{27}$$

where

$$p_{ij} = \frac{e^{\boldsymbol{v}_i^\top\boldsymbol{\beta}_j}}{1 + e^{\boldsymbol{v}_i^\top\boldsymbol{\beta}_j}} \quad \text{and}$$
$$\alpha_{ij} = e^{\boldsymbol{w}_i^\top\boldsymbol{\gamma}_j}, \qquad j = 1,\dots,m; \quad i = 1,\dots,n. \tag{28}$$

Remark 4. The calculation of coefficients usually works well when the dimension is not large. However, the Newton–Raphson algorithm may fail to work when the dimension is high due to the Jacobian (i.e., $\mathbf{J}_{com}(\boldsymbol{\beta})$) tending to be 0 in some iterations. Therefore, studies with a large number of covariates should be carefully handled in order to get reliable estimates. This will be an interesting and practical topic for future research.

5 | HYPOTHESIS TEST

We are usually interested in whether some of the coefficients/parameters are equal to zero. In this section, we will consider the *likelihood ratio test* (LRT) for the following hypotheses:

$$H_0\colon \boldsymbol{\beta}_{i1} = \cdots = \boldsymbol{\beta}_{ir} = \boldsymbol{\gamma}_{j1} = \cdots = \boldsymbol{\gamma}_{jt} = \mathbf{0} \quad \text{against} \quad H_1\colon \text{not } H_0, \tag{29}$$

where $i_1,\dots,i_r$ satisfy $1 \le i_1 < \cdots < i_r \le L_1$, $1 \le j_1 < \cdots < jt \le L_2$, and L_1 and L_2 are the number of covariates related to $\mathbf{p}$ and $\boldsymbol{\alpha}$, respectively. The LRT statistic is then given by

$$T = -2[\ell(\hat{\boldsymbol{\beta}}_0,\hat{\boldsymbol{\gamma}}_0|Y_{\rm obs}) - \ell(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}|Y_{\rm obs})], \tag{30}$$

TABLE 1 MLEs and bootstrap confidence intervals of parameters when $m = 2$

	$n = 100$			$n = 300$			$n = 500$		
True value	**MLE**	**Width**	**CP**	**MLE**	**Width**	**CP**	**MLE**	**Width**	**CP**
$p_1 = 0.1$	0.101	0.128	0.940	0.101	0.075	0.950	0.100	0.058	0.955
$p_2 = 0.2$	0.199	0.163	0.939	0.199	0.094	0.955	0.199	0.073	0.958
$\alpha_1 = 3$	3.131	2.072	0.916	3.049	1.119	0.937	3.024	0.852	0.952
$\alpha_2 = 4$	4.191	2.823	0.910	4.068	1.516	0.930	4.032	1.157	0.946
$p_1 = 0.1$	0.099	0.145	0.935	0.100	0.085	0.935	0.100	0.067	0.950
$p_2 = 0.4$	0.398	0.198	0.960	0.399	0.115	0.941	0.399	0.089	0.945
$\alpha_1 = 2$	2.136	1.711	0.890	2.060	0.889	0.918	2.040	0.682	0.937
$\alpha_2 = 1$	1.053	0.748	0.896	1.025	0.399	0.919	1.017	0.304	0.938

Note: MLE is the mean of the 1000 point estimates via the EM algorithm; width and CP are the average width and coverage proportion of 1000 bootstrap confidence intervals.

where $(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\gamma}}_0)$ are the constrained MLEs of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ under H_0 and $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}$ are the unconstrained MLE of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$. Under the null hypothesis H_0, the p-value is given by

$$p = \Pr(T > t | H_0') = \Pr(\chi^2(\nu) > t), \tag{31}$$

where t is the observed value of T and $\chi^2(\nu)$ is the chi-square distribution with $\nu = r + t$ degrees of freedom.

6 | SIMULATION STUDIES

To evaluate the performance of the proposed statistical methods of DCD, we first investigate the accuracy of point estimates and confidence interval estimates for different parameter settings via simulation studies. We then conduct simulation studies for the regression model. The MLEs of parameters, standard deviation, and confidence intervals are presented. We will compare the ZIGDM model proposed by Tang and Chen (2019) with our proposed DCD model. Finally, simulation results for hypothesis testing are presented. In this section, all statistical computations are implemented in R.

6.1 | Accuracy of point and interval estimates

For the m-dimensional compositional data, there are $2m$ parameters in the DCD (i.e., the m-dimensional parameter $\boldsymbol{p} = (p_1, \dots p_m)^\top$ and m-dimensional parameter $\boldsymbol{\alpha} = (\alpha_1, \dots \alpha_m)^\top$).

We consider two cases, $m = 2$ and $m = 3$, to evaluate the accuracy of point and confidence interval estimates. When $m = 2$, we set $(\boldsymbol{p}, \boldsymbol{\alpha}) = (0.1, 0.2, 3, 4)$ or (0.1,0.4,2,1). When $m = 3$, we set $(\boldsymbol{p}, \boldsymbol{\alpha}) = (0.1, 0.3, 0.2, 3, 2, 4)$ or (0.2,0.2,0.3,2,1,3). For each parameter configuration, we generate $\{\boldsymbol{y}_i\}_{i=1}^n \sim \text{DCD}(\boldsymbol{p}, \boldsymbol{\alpha})$ with $n = 100, 300, 500$, and calculate the MLEs via the EM algorithm and the 95% bootstrap CIs with a significance level $\alpha = 0.05$ with $B = 1000$. The MLEs of parameters, the width, and coverage probability of the bootstrap confidence interval are presented in Tables 1 and 2 for $m = 2$ and $m = 3$, respectively.

From Tables 1 and 2, it is clear that the performance of the MLEs is satisfactory in the sense that (i) the bias of the estimate is negligible; (ii) the confidence width is acceptable; and (iii) the coverage probability is from 0.923 to 0.966, which is not far from the prespecified value $1 - 0.05 = 0.95$. Though 0.923 is a little far from 0.95, the coverage proportion can be improved by increasing the sample size.

6.2 | Numerical results for the regression model

In this subsection, we conduct simulation to evaluate the performance of the proposed regression model for target parameters. Here, we set $m = 3$ and the regression coefficient vector is $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3)$ with the true values being set and

TABLE 2 MLEs and bootstrap confidence intervals of parameters when $m = 3$

	$n = 100$			$n = 300$			$n = 500$		
True value	**MLE**	**Width**	**CP**	**MLE**	**Width**	**CP**	**MLE**	**Width**	**CP**
$p_1 = 0.1$	0.100	0.118	0.936	0.100	0.069	0.940	0.100	0.054	0.953
$p_2 = 0.3$	0.302	0.180	0.950	0.300	0.104	0.943	0.300	0.081	0.950
$p_3 = 0.2$	0.198	0.157	0.943	0.199	0.091	0.938	0.199	0.071	0.938
$\alpha_1 = 3$	3.091	1.528	0.926	3.019	0.840	0.952	3.011	0.646	0.948
$\alpha_2 = 2$	2.046	1.033	0.915	2.010	0.572	0.950	2.007	0.440	0.940
$\alpha_3 = 4$	4.115	2.069	0.924	4.028	1.143	0.943	4.012	0.879	0.939
$p_1 = 0.2$	0.201	0.160	0.953	0.200	0.093	0.948	0.200	0.072	0.955
$p_2 = 0.2$	0.198	0.158	0.953	0.199	0.092	0.959	0.199	0.072	0.949
$p_3 = 0.3$	0.299	0.180	0.954	0.300	0.105	0.948	0.300	0.081	0.952
$\alpha_1 = 2$	2.067	1.088	0.921	2.020	0.595	0.941	2.012	0.456	0.957
$\alpha_2 = 1$	1.031	0.514	0.912	1.012	0.284	0.930	1.007	0.218	0.934
$\alpha_3 = 3$	3.100	1.696	0.925	3.032	0.930	0.939	3.019	0.710	0.945

Note: MLE is the mean of the 1000 point estimates via the EM algorithm; width and CP are the average width and coverage proportion of 1000 bootstrap confidence intervals.

TABLE 3 MLEs for the regression coefficients in the DCD regression model

Parameter	**True**	**MLE**	**Width**	**CP**	**True**	**MLE**	**Width**	**CP**
	0.2	0.211	0.587	0.946	−1	−1.019	0.579	0.915
$\boldsymbol{\beta}_1$	−2	−2.022	0.788	0.950	2	2.030	0.709	0.901
	1	1.013	0.594	0.928	−1	−1.012	0.600	0.972
	1	1.019	0.581	0.967	3	3.063	1.244	0.967
$\boldsymbol{\beta}_2$	−1	−1.016	0.603	0.939	1	1.029	0.713	0.958
	2	2.031	0.700	0.953	−2	−2.037	0.971	0.944
	−1	−1.019	0.721	0.960	−1	−1.016	0.773	0.923
$\boldsymbol{\beta}_3$	−3	−3.059	1.014	0.912	−2	−2.035	1.019	0.918
	3	3.059	1.090	0.936	3	3.047	1.313	0.951
	−1	−0.925	0.357	0.878	−1	−0.986	0.364	0.928
$\boldsymbol{\gamma}_1$	−1	−1.012	0.335	0.924	1	0.971	0.338	0.895
	2	1.975	0.291	0.886	−2	−1.963	0.385	0.919
	−2.5	−2.391	0.312	0.862	−2	−1.992	0.345	0.943
$\boldsymbol{\gamma}_2$	2	1.894	0.316	0.946	0.5	0.478	0.358	0.922
	−2	−1.888	0.386	0.872	−1.5	−1.472	0.358	0.947
	−1	−0.892	0.377	0.939	−1	−0.993	0.382	0.947
$\boldsymbol{\gamma}_3$	1	0.895	0.373	0.959	2	1.974	0.356	0.936
	−2	−1.887	0.341	0.941	−1	−0.973	0.358	0.956

Note: MLE is the mean of the 1000 point estimates via the EM algorithm; width is the mean of the width of the 1000 CIs and CP is the coverage proportion of the confidence intervals.

reported in Table 3. We generate $\{\boldsymbol{x}_i\}_{i=1}^{500} \sim \mathrm{DCD}(\boldsymbol{p}, \boldsymbol{\alpha})$, where $p_{ij} = \frac{\mathrm{e}^{\boldsymbol{v}_i^\top \boldsymbol{\beta}_j}}{1+\mathrm{e}^{\boldsymbol{v}_i^\top \boldsymbol{\beta}_j}}$ and $\alpha_{ij} = \mathrm{e}^{\boldsymbol{w}_i^\top \boldsymbol{\beta}_j}$. For each observed data $\boldsymbol{y}_i$, we calculate the MLEs $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\beta}}_3, \hat{\boldsymbol{\gamma}}_1, \hat{\boldsymbol{\gamma}}_2, \hat{\boldsymbol{\gamma}}_3)$ and this process is repeated 1000 times. The mean value, standard deviation and bootstrap confidence interval are presented in Table 3. According to the results, the MLEs and bootstrap confidence intervals perform well.

TABLE 4 The L_2 distances between observations and predictions: d_1 and d_2

		$\boldsymbol{\pi}_1$		$\boldsymbol{\pi}_2$		$\boldsymbol{\pi}_3$		$\boldsymbol{\pi}_4$	
Parameters		$\boldsymbol{d}_1$	$\boldsymbol{d}_2$	$\boldsymbol{d}_1$	$\boldsymbol{d}_2$	$\boldsymbol{d}_1$	$\boldsymbol{d}_2$	$\boldsymbol{d}_1$	$\boldsymbol{d}_2$
$(\mathbf{a}_1, \mathbf{b}_1)$	n=100	20.26	20.41	21.50	22.06	20.50	20.40	25.57	27.01
	n=300	61.41	61.44	65.31	66.60	62.84	61.61	77.08	81.43
	n=500	102.76	102.34	108.60	110.96	104.80	103.00	128.87	135.42
$(\mathbf{a}_2, \mathbf{b}_2)$	n=100	16.60	16.11	26.66	26.37	36.53	34.79	31.68	32.87
	n=300	49.94	48.44	81.30	79.34	110.94	105.00	96.05	99.18
	n=500	83.07	80.69	135.23	131.94	184.80	175.47	159.84	165.67
$(\mathbf{a}_3, \mathbf{b}_3)$	n=100	14.18	13.79	26.80	26.78	39.48	38.82	32.54	34.27
	n=300	42.66	41.56	81.20	81.19	119.85	117.52	98.16	103.76
	n=500	71.06	69.04	134.98	135.03	200.03	196.08	163.54	172.35

TABLE 5 The L_2 distances between observations and predictions: d_3 and d_4

		$\mathbf{p}_1$		$\mathbf{p}_2$		$\mathbf{p}_3$		$\mathbf{p}_4$	
Parameters		$\boldsymbol{d}_3$	$\boldsymbol{d}_4$	$\boldsymbol{d}_3$	$\boldsymbol{d}_4$	$\boldsymbol{d}_3$	$\boldsymbol{d}_4$	$\boldsymbol{d}_3$	$\boldsymbol{d}_4$
$\boldsymbol{\alpha}_1$	n=100	40.86	38.72	26.08	21.35	28.90	26.17	30.54	29.37
	n=300	122.35	116.52	78.82	64.76	87.35	79.20	92.61	89.08
	n=500	204.86	195.15	131.57	107.93	145.43	131.91	154.36	148.83
$\boldsymbol{\alpha}_2$	n=100	56.19	52.03	23.50	21.80	36.75	34.89	50.54	50.60
	n=300	169.76	156.88	70.18	65.42	110.42	104.67	152.10	152.17
	n=500	283.58	261.86	117.68	109.50	184.64	174.48	254.46	254.46
$\boldsymbol{\alpha}_3$	n=100	52.21	50.98	37.92	31.52	44.45	41.45	34.01	31.12
	n=300	157.61	153.92	113.95	94.95	132.90	124.44	103.05	94.36
	n=500	263.51	257.16	190.19	158.34	223.38	207.44	172.41	157.27

6.3 | The sensitivity and robustness of the model

In this subsection, we will investigate the sensitivity and robustness of our model and compare the performance between the DCD and ZIGDM models. Let $\mathbf{u} = (u_1, \ldots, u_m)^\top$ be the observation in the ZIGDM model. It is easy to see that the MLEs for the parameters based on $\mathbf{u}$ and $\mathbf{u}' = c\mathbf{u}$ with c being any nonzero constant are different under the ZIGDM model. However, the MLE for the parameters under our proposed DCD model will be invariant for constant multiplication. From this perspective, our model is more robust than the ZIGDM model. Next, we will consider the sensitivity of our model. For this purpose, we generate the data from the ZIGDM model and compare the L_2 distance between the observation and the predictions using the DCD model and ZIGDM model. We set sample size $n = 100, 300, 500$ and parameters being $\boldsymbol{\pi}_1 = (0, 0.1)^\top$, $\boldsymbol{\pi}_2 = (0.1, 0.1)^\top$, $\boldsymbol{\pi}_3 = (0.2, 0)^\top$, $\boldsymbol{\pi}_4 = (0.2, 0.3)^\top$; $\mathbf{a}_1 = (2, 3)^\top$, $\mathbf{b}_1 = (5, 3)^\top$, $\mathbf{a}_2 = (3, 3)^\top$, $\mathbf{b}_2 = (2, 4)^\top$, $\mathbf{a}_3 = (4, 1)^\top$, and $\mathbf{b}_3 = (2, 2)^\top$. We generate the data $\{\mathbf{u}_i\}_{i=1}^n$ with $\mathbf{u}_i = (u_{i1}, \ldots, u_{im})^\top$ and each $\mathbf{u}_i$ following the ZIGDM distribution with $N = N_1 = \cdots = N_n = 100$. That is $\mathbf{u}_i \sim \text{ZIGDM}(\boldsymbol{\pi}, \mathbf{a}, \mathbf{b})$. For $\{\mathbf{u}_i\}_{i=1}^n$, we obtain the MLEs of the parameters $(\hat{\boldsymbol{\pi}}, \hat{\mathbf{a}}, \hat{\mathbf{b}})$, and then calculate the predictions of the $\{\hat{\mathbf{u}}_i^{(1)}\}_{i=1}^n$ by applying the ZIGDM model. Similarly, the predictions of $\{\hat{\mathbf{u}}_i^{(2)}\}_{i=1}^n$ are obtained based on the DCD model. The L_2 distance between observations and predictions (denoted as d_1 and d_2) are presented in Table 4, where $d_k = \sum_{i=1}^n (\mathbf{u}_i - \hat{\mathbf{u}}_i^{(k)})^\top (\mathbf{u}_i - \hat{\mathbf{u}}_i^{(k)})/N^2$ with $k = 1$ and 2.

As the data are generated from the DM model, d_1 is expected to be less than d_2. From Table 4, d_1 is generally less than d_2. It is noticed that the DCD model sometimes fits better than the ZIDGM model. It suggests that our proposed DCD model is robust.

Next, we generate the data from DCD distribution with $n = 100, 300, 500$ and parameters being $\mathbf{p}_1 = (0.4, 0.1, 0.3)^\top$, $\mathbf{p}_2 = (0, 0.1, 0.2)^\top$, $\mathbf{p}_3 = (0.2, 0, 0.3)^\top$, $\mathbf{p}_4 = (0.3, 0.2, 0.1)^\top$; $\boldsymbol{\alpha}_1 = (4, 5, 6)^\top$, $\boldsymbol{\alpha}_2 = (3, 2, 1),^\top$ and $\boldsymbol{\alpha}_3 = (2, 2, 5)^\top$. We obtain the MLEs based on the ZIGDM model and then calculate the L_2 distance between the predictions and observed data (denoted as d_3). Since the ZIGDM model does not work for DCD data, we assume $N_1 = \cdots = N_m = 100$. Similarly, we can get the L_2 distance (denoted as d_4) using the DCD model. We report d_3 and d_4 in Table 5. As expected, d_3 should be generally larger than d_4. The results in Table 5 support that the performance of the DCD model is generally better than the ZIGDM model.

TABLE 6 The type I error rates for the LRT

Type I error	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
Case I	0.057	0.073	0.057	0.045	0.056
Case II	0.100	0.058	0.040	0.053	0.050
Case III	0.077	0.056	0.047	0.047	0.044

TABLE 7 The simulated powers for the LRT

Type I error	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
Case I	0.692	0.955	1.000	1.000	1.000
Case II	1.000	1.000	1.000	1.000	1.000
Case III	1.000	1.000	1.000	1.000	1.000

6.4 | Hypothesis testing

In this subsection, we evaluate the performance of our proposed LRT. We set $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2}, \beta_{j3})^\top$ and $\boldsymbol{\gamma}_j = (\gamma_{j1}, \gamma_{j2}, \gamma_{j3})^\top$ for $j = 1, 2, 3$. According to Equation (8), we first consider Case I:

$$H_0\colon \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}_3 = 0 \quad \text{against} \quad H_1\colon H_0 \text{ is not true.} \tag{32}$$

We generate the data $\{\boldsymbol{x}_i\}_{i=1}^n \sim \text{DCD}(\boldsymbol{p}, \boldsymbol{\alpha})$ for 1000 times, where $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}_3 = \boldsymbol{0}$ and $\boldsymbol{\gamma}_1 = (-1, -1, 2)^\top$, $\boldsymbol{\gamma}_2 = (-2.5, 2, -2)^\top$, and $\boldsymbol{\gamma}_3 = (-1, 1, -2)^\top$. Second, we consider Case II:

$$H_0\colon \boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2 = \boldsymbol{\gamma}_3 = \boldsymbol{0} \quad \text{against} \quad H_1\colon H_0 \text{ is not true.} \tag{33}$$

We generate the data with $\boldsymbol{\gamma}_1' = \boldsymbol{\gamma}_2' = \boldsymbol{\gamma}_3' = \boldsymbol{0}$ and $\boldsymbol{\beta}_1' = (0.2, -2, 1)^\top$, $\boldsymbol{\beta}_2' = (0.3, -1, 0.5)^\top$, and $\boldsymbol{\beta}_3' = (-1, -3, 3)^\top$. Third, we consider Case III:

$$H_0\colon \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}_3 = \boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2 = \boldsymbol{\gamma}_3 = 0 \quad \text{against} \quad H_1\colon H_0 \text{ is not true.} \tag{34}$$

We generate data with parameters being $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}_3 = \boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2 = \boldsymbol{\gamma}_3 = \boldsymbol{0}$. Applying the LRT in Section 5, we record the proportions of rejection of the above three cases with the sample size being $n = 50$, $n = 100$, $n = 200$, $n = 500$, $n = 1000$. The simulated type I error rate is reported in Table 6. It is clearly that the performance of our LRT is fairly good even when the sample size is small. When the sample size is larger than 200, the simulated type I error rate is close to the prespecified significance level (i.e., 0.05).

Next, we investigate the power of the LRT. We generate the data from $\{\boldsymbol{x}_i\}_{i=1}^n \sim \text{DCD}(\boldsymbol{p}, \boldsymbol{\alpha})$ with parameters being $(\boldsymbol{0}, \boldsymbol{\beta}_2', \boldsymbol{0}, \boldsymbol{0}, \boldsymbol{\gamma}_2, \boldsymbol{0})$. The number of rejections according to the above three cases is recorded in Table 7. As expected, the simulated power increases with the sample size.

7 | THE PERCENTAGE OF CHROMOSOME DATA BY THE FISH TEST

In this section, we revisit the FISH test dataset described in Section 1. We here apply the DCD to analyze the FISH data of chromosome 16. First, we analyze the dataset with no covariate, and the results are reported in Table 8.

As we can see from Table 8, the first component in the composition dataset (i.e., the normal cell) is all nonzeros, $\hat{p}_1 = 0$ means that the probability of zero observation for the normal cell is 0. The probability of zero observations of the triple and tetraploid cell is estimated to be 0.784 and 0.980, respectively. That is, for the chromosome 16 the triple cell can be found with 20% percentage while tetraploid can be found with only 2% percentage. For the base part of the compositional data, the estimate corresponding to the triple cell is larger than those of the normal and tetraploid cells; that is, $\hat{\alpha}_1 = 12.17$, $\hat{\alpha}_1 = 38.52$, and $\hat{\alpha}_3 = 6.12$. In other words, once there is abnormal chromosome in the cell the triple is more frequent than the other two.

TABLE 8 MLEs and 95% bootstrap CIs of parameters for the parameters of the FISH data

Parameter	MLE	Mean	Median	95% Bootstrap CI
p_1	0.000	0.000	0.000	[0.000, 0.000]
p_2	0.784	0.774	0.765	[0.667, 0.882]
p_3	0.980	0.969	0.980	[0.922, 0.980]
α_1	12.170	10.860	10.435	[6.661, 17.288]
α_2	38.516	33.009	32.074	[17.123, 54.360]
α_3	6.117	5.359	5.202	[3.146, 8.399]

TABLE 9 MLEs and 95% bootstrap CIs of parameters for the coefficients in the regression model of the FISH data

Parameter	Coefficients	|cMLE	Standard deviation	95% Bootstrap CI
β_1	Intercept	−12.985	2.912	[−19.302, −8.034]
	Age	1.019	1.840	[−2.419, 4.221]
β_2	Intercept	1.293	0.371	[0.663, 2.076]
	Age	0.080	0.388	[−0.680, 0.942]
β_3	Intercept	4.192	0.459	[2.779, 4.448]
	Age	0.796	0.316	[0.458, 1.753]
γ_1	Intercept	3.116	0.559	[1.238, 3.798]
	Age	1.003	0.992	[−2.461, 2.082]
γ_2	Intercept	4.270	0.652	[1.876, 4.904]
	Age	1.178	1.036	[−2.534, 2.234]
γ_3	Intercept	1.234	0.329	[0.729, 1.803]
	Age	− 0.615	0.150	[−0.925, −0.366]

Due to the assumption of the normal distribution for covariates in the regression model, we make the standardization for covariate age in the FISH dataset. Furthermore, we apply the regression model to investigate the relationship between parameters $(\boldsymbol{p}, \boldsymbol{\alpha})$ and the age of gravida. The MLEs of the coefficients are reported in Table 9. We apply the LRT to the hypotheses listed in Section 5, and the null hypothesis is rejected at $\alpha = 0.05$. Therefore, we have reason to believe that age is a significant variable.

8 | DISCUSSION

In this article, we consider a new framework for analyzing compositional data with zero entries based on SR. In particular, a new distribution, namely the DCD, is developed to accommodate the possible essential-zero feature in compositional data (i.e., some components are completely absent). In our proposed distribution, the elements in the base vector are assumed to follow independent gamma distributions. Therefore, any positive random variable can be adopted as an element of the base vector (e.g., the inverse Gaussian and chi-square random variables), and different base vectors will correspond to a different relationship among elements. It is of research and practical interests to relax the assumption of independence among the components in the base vector. Besides, regression modeling for high-dimensional covariates is also an interesting and necessary topic as the Jacobi tends to be zero when the dimension of covariates is high.

ACKNOWLEDGMENTS

The research of Qin Wu was supported by the National Natural Science Foundation of China (No. 12171167 and 11801184). The work of Man-Lai Tang was supported through grants from the Research Grant Council of the Hong Kong Special Administrative Region (projects UGC/FDS14/P06/17 and UGC/FDS14/P02/18) and a grant from the National Natural Science Foundation of China (grant 11871124). Guo-Liang Tian's research was partially supported by the National Natural Science Foundation of China (No. 11771199 and 12171225).

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

Data are available on request from the authors.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to their computational complexity.

ORCID

Man-Lai Tang https://orcid.org/0000-0003-3934-2676
Qin Wu https://orcid.org/0000-0002-9986-7530

REFERENCES

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B*, *44*(2), 139–177.

Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman & Hall.

Aitchison, J., & Kay, J. W. (1990). Possible solution of some essential zero problems in compositional data analysis. Paper presented at CODA-WORK03. http://links.jstor.org/sici?sici=0035-9246%281982%2944%3A2%3C139%3ATSAOCD%3E2.0.CO%3B2-9.

Bear, I. J., & Billheimer, D. (2016). A logistic normal mixture model allowing essential zeros. *Austrian Journal of Statistics*, *45*, 3–23.

Butler, A., & Glasbey, C. (2008). A latent Gaussian model for compositional data with zeros. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *57*(5), 505–520.

Egozue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., & Barceló-Vidal, C. (2003). Isometric log-ratio transformations for compositional data analysis. *Mathematical Geology*, *35*(3), 279–300.

Ferrers, N. M. (1866). *An elementary treatise on trilinear coordinates*. Macmillan.

Hijazi, R. H. (2011). An EM-algorithm based method to deal with rounded zeros in compositional data under Dirichlet models. Paper presented at the Proceedings of the First Compositional Data Analysis Workshop, Girona, Spain

Kent, J. T. (1982). The Fisher–Bingham distribution on the sphere. *Journal of the Royal Statistical Society. Series B (Methodological)*, *44*(1), 71–80.

Langer-Safer, P. R., Levine, M., & Ward, D. C., (1982). Immunological method for mapping genes on Drosophila polytene chromosomes. *Proceedings of the National Academy of Sciences*, *79*(14), 4381–4385.

Martín-Fernández, J. A., Barceló-Vidal, C., & Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, *35*, 253–278.

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, *49*, 65–82.

Pearson, K. (1897). Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, *60*, 489–502.

Palarea-Albaladejo, J., & Martín-Fernández, J. A. (2008). A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. *Computers & Geosciences*, *34*, 902–917.

Scealy, J., & Welsh, A. (2011). Regression for compositional data by using distributions defined on the hypersphere. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(3), 398–412.

Stewart, C., & Field, C. (2011). Managing the essential zeros in quantitative fatty acid signature analysis. *Journal of Agricultural, Biological, and Environmental Statistics*, *16*(1), 45–69.

Tang, Z. Z., & Chen, G. (2019). Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics*, *20*(4), 698–713

Tsagris, M., & Stewart, C. (2018). A Dirichlet regression model for compositional data with zeros. *Lobachevskii Journal of Mathematics*, *39*(3), 398–412.

Zadora, G., Neocleous, T., & Aitken, C. (2010). A two-level model for evidence evaluation in the presence of zeros. *Journal of Forensic Sciences*, *55*, 371–384.

Zhang, R. T. (2000). *An introduction to compositional data analysis*. Chinese Science Press.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Tang, M.-L., Wu, Q., Yang, S., & Tian, G.-L. (2021). Dirichlet composition distribution for compositional data with zero components: An application to fluorescence in situ hybridization (FISH) detection of chromosome. *Biometrical Journal*. 1–19. https://doi.org/10.1002/bimj.202000334

APPENDIX

Before we present the proofs, we first introduce the following two lemmas.

Lemma 1. *Let $\mathbf{X} = (X_1, \ldots, X_m)^\top$ follow the composition Dirichlet distribution DCD($\boldsymbol{p}, \boldsymbol{\alpha}$), $\mathbf{Y}$ be the base vector with $Y_i \overset{\text{ind}}{\sim} \Gamma(\alpha_i, 1)$ and $\mathbf{Z}$ be an indicator vector with $\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}$. If $s = \sum_{i=1}^{m} Z_i Y_i$, then the joint probability density function of $(\mathbf{X}, s)$ conditioned on $\mathbf{Z}$ is given by*

$$f(\boldsymbol{x}, s|\boldsymbol{z}) = f(\boldsymbol{x}_{\mathbb{J}}, s) = s^{\alpha-1} e^{-s} \times \left(\prod_{j \in \mathbb{J}} \frac{x_j^{\alpha_j - 1}}{\Gamma(\alpha_j)} \right), \tag{A.1}$$

where $\mathbb{J}$ is the subset of the index with $\mathbf{X}$ being positive; that is, we have $x_j > 0$ for $j \in \mathbb{J}$, $x_j = 0$ for $j \notin \mathbb{J}$ and $\sum_{j=1}^{m} X_j = 1$.

Proof. Without loss of generality, we assume that the first r elements of $\mathbf{X}$ is positive; that is, $\mathbb{J} = \{1, 2, \ldots, r\}$. The joint probability density function of $(Y_1, \ldots, Y_r)$ is then given by

$$f(y_1, \ldots, y_r) = \prod_{j=1}^{r} \frac{y_j^{\alpha_j - 1} \mathrm{e}^{-y_j}}{\Gamma(\alpha_j)}. \tag{A.2}$$

Next, we consider the following transformation:

$$\begin{cases} s & = & \sum_{i=1}^{r} Y_i. \\ X_1 & = & Y_1/s, \\ & \vdots & \\ X_{r-1} & = & Y_{r-1}/s. \end{cases} \Longleftrightarrow \begin{cases} Y_1 & = & sX_1, \\ & \vdots & \\ Y_{r-1} & = & sX_{r-1}, \\ Y_r & = & s(1 - \sum_{i=1}^{r-1} X_i). \end{cases} \tag{A.3}$$

Obviously, $\mathbf{Y}_r = (Y_1, \ldots, Y_r)^\top$ and $(s, \mathbf{X}_{r-1})^\top = (s, X_1, \ldots, X_{r-1})^\top$ are one-to-one transformations. The Jacobian determinant from $\mathbf{Y}_r$ to $(s, \mathbf{X}_{r-1})$ is given by

$$J(\mathbf{y}_r | s, \mathbf{X}_{r-1}) = \begin{vmatrix} X_1 & s & 0 & \cdots & 0 \\ X_2 & 0 & s & \cdots & 0 \\ \vdots & 0 & 0 & \ddots & 0 \\ X_{r-1} & 0 & 0 & \cdots & s \\ 1 - \sum_{i=1}^{r-1} X_i & -s & -s & \cdots & -s \end{vmatrix}$$

$$= \begin{vmatrix} X_1 & s & 0 & \cdots & 0 \\ X_2 & 0 & s & \cdots & 0 \\ \vdots & 0 & 0 & \ddots & 0 \\ X_{r-1} & 0 & 0 & \cdots & s \\ 1 & 0 & 0 & \cdots & 0 \end{vmatrix} = s^{r-1}. \tag{A.4}$$

The probability density function of $(s, \mathbf{X}_{r-1})$ given $\mathbf{Z}$ is then given by

$$f(s,\boldsymbol{x}_{r-1}|\boldsymbol{z}) = f(\mathbf{y}_r)J(\mathbf{y}_r|s,\boldsymbol{x}_{r-1}) = s^{r-1}\prod_{j=1}^{r}\frac{(sx_j)^{\alpha_j-1}\mathrm{e}^{-sx_j}}{\Gamma(\alpha_j)} = s^{\alpha^*_{\mathbb{J}}-1}\mathrm{e}^{-s}\prod_{j=1}^{r}\frac{x_j^{\alpha_j-1}}{\Gamma(\alpha_j)}, \tag{A.5}$$

where $\alpha^*_{\mathbb{J}} = \sum_{j=1}^{r}\alpha_j$, $x_r = 1 - \sum_{i=1}^{r-1} x_i$. □

Lemma 2. *Let* $\mathbf{X} = (X_1, \dots, X_m)^\top$ *follow the composition Dirichlet distribution DCD(*$\boldsymbol{p}, \boldsymbol{\alpha}$*),* $\mathbf{Y}$ *be the base vector with* $Y_i \overset{\text{ind}}{\sim} \Gamma(\alpha_i, 1)$ *and* $\mathbf{Z}$ *be an indicator vector with* $\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}$*. If* $s = \sum_{i=1}^{m} Z_i Y_i$*, then the probability density function of* $\mathbf{X}$ *conditioned on* $\mathbf{Z}$ *is given by*

$$f(\boldsymbol{x}|\boldsymbol{z}) = f(\boldsymbol{x}_{\mathbb{J}}) = \Gamma(\alpha^*_{\mathbb{J}})\prod_{j\in\mathbb{J}}\frac{x_j^{\alpha_j-1}}{\Gamma(\alpha_j)} = \Gamma(\alpha^*_{\mathbb{J}})\prod_{j\in\mathbb{J}}\frac{x_j^{\alpha_j-1}}{\Gamma(\alpha_j)}, \tag{A.6}$$

where $\alpha^*_{\mathbb{J}} = \sum_{j=1}^{r}\alpha_j$.

Proof.

$$f(\boldsymbol{x}_{\mathbb{J}}) = \int_0^{+\infty} f(s, \boldsymbol{x}_{\mathbb{J}})ds = \int_0^{+\infty} s^{\alpha^*_{\mathbb{J}}-1}\mathrm{e}^{-s}\prod_{j\in\mathbb{J}}\frac{x_j^{\alpha_j-1}}{\Gamma(\alpha_j)}ds = \Gamma(\alpha^*_{\mathbb{J}})\prod_{j\in\mathbb{J}}\frac{x_j^{\alpha_j-1}}{\Gamma(\alpha_j)}. \tag{A.7}$$

□

A.1 | Proof of the pdf of DCD

The probability density function of the DCD is

$$f(\boldsymbol{x}) = \frac{\prod_{j=1}^{m} p_j^{1-z_j}(1-p_j)^{z_j}}{1 - p_1 \cdots p_m} \times \left[\frac{\Gamma(\alpha^*_{\mathbb{J}_i})}{\prod_{j\in\mathbb{J}}\Gamma(\alpha_j)}\prod_{j\in\mathbb{J}} x_j^{\alpha_j-1}\right]. \tag{A.8}$$

Proof. The pdf of the DCD can be derived via its SR; that is, $\mathbf{X} \overset{\mathrm{d}}{=} \frac{\mathbf{Z}\circ\mathbf{Y}}{(\mathbf{Z},\mathbf{Y})}$. Actually, the indicator vector $\boldsymbol{z}$ is derived from $\boldsymbol{x}$ by $z_j = I(x_j > 0)$, $j = 1, \dots, m$. It is clear that the indicator vector $\mathbf{Z}$ is determined by the compositional data $\mathbf{X}$, the conditional mass function of $\mathbf{Z}$ is $\Pr(Z_1 = z_1, \dots, Z_m = z_m|\boldsymbol{x}) = I(z_1 = I(x_1 > 0), \dots, z_m = I(z_m > 0))$. Thus we have $f(\boldsymbol{x}, \boldsymbol{z}) = f(\boldsymbol{z}|\boldsymbol{x})f(\boldsymbol{x}) = f(\boldsymbol{x})$

Therefore, the probability density function (pdf) of DCD is

$$f(\boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{z}) = f(\boldsymbol{x}|\boldsymbol{z})f(\boldsymbol{z}) = \frac{\Gamma(\alpha^*_{\mathbb{J}_i})}{\prod_{j\in\mathbb{J}}\Gamma(\alpha_j)}\prod_{j\in\mathbb{J}} x_j^{\alpha_j-1} \times \frac{\prod_{j=1}^{m}(1-p_j)^{z_j}p_j^{1-z_j}}{1 - p_1 \cdots p_m}. \tag{A.9}$$

□

A.2 | Proof of Theorem 1

Lemma 3. *Let* $\mathbf{X} = (X_1, \dots, X_m)^\top$ *follow the composition Dirichlet distribution DCD(*$\boldsymbol{p}, \boldsymbol{\alpha}$*),* $\mathbf{Y}$ *be the base vector with* $Y_i \overset{\text{ind}}{\sim} \Gamma(\alpha_i, 1)$ *and* $\mathbf{Z}$ *be an indicator vector with* $\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}$*. If* $s = \sum_{i=1}^{m} Z_i Y_i$*, then the joint probability density function of* $(Y_j, \mathbf{X}_{\mathbb{J}})$ *is given* *as follows:*

$$f(y_j, \boldsymbol{x}_{\mathbb{J}}) = f(s, \boldsymbol{x}_{\mathbb{J}})\frac{1}{x_j} = \frac{1}{x_j} s^{\alpha^*_{\mathbb{J}}-1} e^{-s} \prod_{j\in\mathbb{J}} \frac{x_j^{\alpha_j-1}}{\Gamma(\alpha_j)}, \tag{A.10}$$

where $j \in \mathbb{J}$, $\alpha^*_{\mathbb{J}} = \sum_{j=1}^{r} \alpha_j$, $s = y_j/x_j$.

Proof. The proof follows immediately by using the transformation $s = y_j/x_j$. □

Lemma 4. *Let* $\mathbf{X} = (X_1, \dots, X_m)^\top$ *follow the composition Dirichlet distribution DCD*$(\boldsymbol{p}, \boldsymbol{\alpha})$*,* $\mathbf{Y}$ *be the base vector with* $Y_i \overset{\text{ind}}{\sim} \Gamma(\alpha_i, 1)$ *and* $\mathbf{Z}$ *be an indicator vector with* $\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}$*. If* $s = \sum_{i=1}^{m} Z_i Y_i$*, then we have*

$$E(\log y_j | x_j > 0) = \psi(\alpha^*_{\mathbb{J}_i}) + \log x_j, \tag{A.11}$$

where $\mathbb{J}$ *is the subset of the index with* $\boldsymbol{x}$ *being positive; that is, for any* $j \in \mathbb{J}$*, we have* $x_j > 0$*; otherwise* $x_j = 0$*.*

Proof.

$$\begin{aligned}
E(\log y_j | \boldsymbol{x}) &= \int_0^{+\infty} \log y_j f(y_j|\boldsymbol{x}) dy_j = \int_0^{+\infty} \log y_j \times \frac{\frac{1}{x_j} s^{\alpha^*_{\mathbb{J}}-1} \mathrm{e}^{-s} \prod_{j\in\mathbb{J}} \frac{x_j^{\alpha_j-1}}{\Gamma(\alpha_j)}}{\Gamma(\alpha^*_{\mathbb{J}}) \prod_{j\in\mathbb{J}} \frac{x_j^{\alpha_j-1}}{\Gamma(\alpha_j)}} dy_j \\
&\overset{(2)}{=} \int_0^{+\infty} \frac{(\log s + \log x_j)\mathrm{e}^{-s} s^{\alpha^*_{\mathbb{J}}-1}}{\Gamma(\alpha^*_{\mathbb{J}})} ds \\
&= \int_0^{+\infty} \frac{(\log s) \times \mathrm{e}^{-s} s^{\alpha^*_{\mathbb{J}}-1}}{\Gamma(\alpha^*_{\mathbb{J}})} ds + \int_0^{+\infty} \log x_j \times \frac{\mathrm{e}^{-s} s^{\alpha^*_{\mathbb{J}}-1}}{\Gamma(\alpha^*_{\mathbb{J}})} ds \\
&= \psi(\alpha^*_{\mathbb{J}}) + \log x_j.
\end{aligned} \tag{A.12}$$

Note: We make the transformation $y_j = s x_j$ above. □

Lemma 5. *Let* $\mathbf{X} = (X_1, \dots, X_m)^\top$ *follow the composition Dirichlet distribution DCD*$(\boldsymbol{p}, \boldsymbol{\alpha})$*,* $\mathbf{Y}$ *be the base vector with* $Y_i \overset{\text{ind}}{\sim} \Gamma(\alpha_i, 1)$ *and* $\mathbf{Z}$ *be an indicator vector with* $Z_i \sim$ *Bernoulli*$(1 - p_i)$ *and* $\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}$*. If* $s = \sum_{i=1}^{m} Z_i Y_i$*, then we have*

$$E(\log y_j | x_j = 0) = \psi(\alpha_j). \tag{A.13}$$

Proof.

$$\begin{aligned}
E(\log y_j | x_j = 0) = E(\log y_j | z_j = 0) &= E(\log y_j) \\
&= \int_0^{+\infty} \log y_j f(y_j) dy_j = \int_0^{+\infty} \log y_j \times \frac{y_j^{\alpha_j-1} \mathrm{e}^{-y_j}}{\Gamma(\alpha_j)} dy_j \\
&= \psi(\alpha_j).
\end{aligned} \tag{A.14}$$

Based on the results in Lemmas 4 and 5, Theorem 1 follows immediately with

$$E(\log y_j|\boldsymbol{x}) = \begin{cases} \psi(\alpha_j), & \text{if} \quad x_j = 0, \\ \psi(\alpha^*_{\mathbb{J}_i}) + \log x_j & \text{if} \quad x_{ij} > 0. \end{cases} \tag{A.15}$$

□