Statistical Papers (2025) 66:113
https://doi.org/10.1007/s00362-025-01732-8

REGULAR ARTICLE

On the distribution of isometric log-ratio coordinates under extra-multinomial count data

Noora Kartiosuo[1,2,3,4] · Joni Virta[1] · Jaakko Nevalainen[5] · Olli Raitakari[2,3,6] · Kari Auranen[1,7]

Received: 23 August 2024 / Revised: 28 April 2025

Abstract
Compositional data can be mapped from the simplex to the Euclidean space through the isometric log-ratio (ilr) transformation. When the underlying counts follow a multinomial distribution, the distribution of the ensuing ilr coordinates has been shown to be asymptotically multivariate normal. We derive conditions under which the asymptotic normality of the ilr coordinates holds under a compound multinomial distribution inducing overdispersion in the counts. We derive a normal approximation and investigate its practical applicability under extra-multinomial variation using a simulation study under the Dirichlet-multinomial distribution. The approximation works well, except with a small total count or high amount of overdispersion. Our work is motivated by microbiome data, which exhibit extra-multinomial variation and are increasingly treated as compositions. We conclude that if empirical data analysis relies on the normality of ilr coordinates, it may be advisable to choose a taxonomic level with less sparsity so that the distribution of taxon-specific class probabilities remains unimodal.

Keywords Asymptotic approximation · Compositional data analysis · Dirichlet-multinomial · Isometric log-ratio transformation · Sequencing count data

MSC Classification 62E20

1 Introduction

Compositional data involve elements which together constitute a whole entity so that the measurements only carry relative information. Such data may arise when non-negative observations, such as counts, are considered as compositions through scaling individual observations by their total into proportions that sum up to unity, or

Extended author information available on the last page of the article

Published online: 30 June 2025

more generally, are constrained to add up to a constant. The total sum of the elements is considered irrelevant, while the information carried by the relative measures is of interest. This means that two samples can be compositionally equivalent even if their totals deviate. The proportions can be modelled to lie in the simplex, and due to this constant-sum constraint, traditional statistical methods may not be applicable. For these type of data, the framework of compositional data analysis provides a suitable collection of methods (Aitchison 1982, 1986).

Importantly, the simplex has a Euclidean vector space structure where different coordinate systems can be specified using different log-ratio transformations (Mateu-Figueras et al. 2011; Aitchison 1986). Particularly convenient is the isometric log-ratio transformation (ilr), which maps the simplex equipped with the so-called Aitchison geometry to the standard Euclidean vector space of coordinates with respect to an orthonormal basis (Egozcue et al. 2003). The isometric relationship between the two spaces allows basing probabilistic modelling for vectors within the simplex directly on their ilr coordinates (Mateu-Figueras et al. 2011). In particular, the so-called normal on the simplex is obtained by assuming that the ilr coordinates have a multivariate normal distribution (Mateu-Figueras et al. 2013). The benefit of "working on coordinates" is that standard statistical methods, such as linear regression models, are readily applicable.

Recently, the use of compositional data analysis has been strongly endorsed also in microbiome research, where datasets consist of read counts of a number of distinct taxa, but the total count is arbitrary due to varying sequencing depths and thus does not carry any substantive information (Gloor et al. 2017; Egozcue et al. 2020). Different log-ratio transformations have been applied to microbial counts assuming the normality of log-ratio coordinates (Sohn and Li 2019; Zhang et al. 2021). When the underlying counts follow the multinomial distribution, the induced distribution of the ensuing ilr coordinates, viewed basically as a data transformation, has indeed been shown to be asymptotically multivariate normal (Graffelman 2011; Graffelman et al. 2015). Thus, given a large enough total count, it is justifiable to apply methods based on the assumption of normality of the ilr coordinates. Furthermore, the mean and covariance parameters of the limiting normal distribution of the ilr coordinates can then be expressed in terms of multinomial probability parameters.

Nevertheless, microbial counts are characterised by heterogeneity across individuals in taxon-specific class probabilities, which leads to overdispersion and excess of zeroes in class-specific counts in regard to the simple multinomial distribution. A possible choice for addressing such heterogeneity is the logratio-normal-multinomial distribution, where the class-specific proportions are modelled as inverses of the ilr coordinates (Mateu-Figueras et al. 2013). However, this approach makes the assumption that the ilr coordinates are normally distributed. Alternatively, overdispersed counts can be obtained by treating counts as realisations from a compound multinomial distribution where the class-specific probabilities are not fixed but follow some distribution. The standard choice among such compound multinomial distributions is the Dirichlet-multinomial model. Although sometimes criticised due to imposing negative correlations between parts of compositions (Xia et al. 2013; Comas-Cufí et al. 2020), potentially spurious in the analysis of microbial data, this distribution is widely used (Fernandes et al. 2013; Wang et al. 2020).

At the extreme levels of heterogeneity, the distribution of proportions becomes multimodal, causing sample-specific counts to concentrate within some of the classes and leading to zero counts in many other classes. It is obvious that under such a large overdispersion the distribution of class-specific proportions is multimodal and the asymptotic normality of the induced distribution of the ilr coordinates cannot hold. However, it still remains to be investigated how the amount of extra-multinomial overdispersion and the ensuing sparsity, i.e., excess of zero-count observations with respect to the multinomial model, affect the normality of the ilr coordinates in finite samples. The aim of this paper is to derive a normal approximation for the ilr coordinates and investigate conditions under which the asymptotic normality of the ilr coordinates holds when based on compound multinomial counts, i.e., on count observations that exhibit extra-multinomial variation. Apart from general theoretical results we consider a simulation study where, for convenience and transparency of interpretation, we focus on counts that follow a Dirichlet-multinomial distribution.

The contributions of this work are as follows:

- We present conditions under which the asymptotic normality of compound multinomial counts holds under two alternative conditions for the mixing distribution (Sect. 2).
- Assuming a Dirichlet-multinomial distribution of counts, we apply the results of Sect. 2 and present conditions for the asymptotic normality of counts and the ensuing proportions (Sect. 3).
- Based on the asymptotic normality of the proportions under compound multinomial models as derived in Sects. 2 and 3, we argue for the asymptotic normality of the induced ilr coordinates and derive an explicit normal approximation under the special case of the Dirichlet-multinomial model (Sect. 4).
- Using a simulation study, we investigate how well the asymptotic normality and the moments of the normal approximation hold under the Dirichlet-multinomial model in the presence of varying extents of extra-multinomial variation (Sect. 5).
- Additionally, we study how variability in the sample-specific total count influences the distribution of the ilr coordinates (Sects. 4 and 5).
- Finally, in Sect. 6 we summarise our findings and discuss their implications on the analysis of microbiome data.

2 Asymptotic normality of compound multinomial distributions

Denote the J-part unit simplex by $\mathcal{S}^{J-1}$. We here consider the asymptotic behaviour of a compound multinomial variable $\boldsymbol{x}_K$, $\boldsymbol{x}_K \mid \boldsymbol{\pi}_K \sim \text{Multinomial}(K, \boldsymbol{\pi}_K)$ where the probability parameter $\boldsymbol{\pi}_K \in \mathcal{S}^{J-1}$ admits a decomposition $\boldsymbol{\pi}_K = \boldsymbol{\alpha} + \boldsymbol{z}_K$, where $\boldsymbol{\alpha} \in \mathcal{S}^{J-1}$ is fixed and $\boldsymbol{z}_K$ is such a random vector in $\mathbf{R}^J$ that the sum $\boldsymbol{\alpha} + \boldsymbol{z}_K$ stays on the simplex $\mathcal{S}^{J-1}$.

The distribution of $\boldsymbol{z}_K$ determines the mixing distribution for the vector of multinomial probabilities. The index K corresponds to the total count of the observation $\boldsymbol{x}_K$ and we present our results in the asymptotic scenario where $K \to \infty$. This regime approximates the practical scenario where the total observed count is suf-

ficiently large. Note that throughout this paper we consider the asymptotic distributions of *observations* (counts and their proportions) and the ensuing ilr coordinates, rather than the asymptotic behaviour of parameter estimators.

The following theorems are based on two alternative conditions. The first condition requires fast enough convergence of $\boldsymbol{z}_K$ to 0 as $K \to \infty$. The role of the mixing distribution is then asymptotically negligible and the asymptotic normality of $\boldsymbol{x}_K$ follows from the normal limiting distribution of the multinomial distribution (Theorem 1). The second condition assumes the existence of a limiting distribution for an appropriately scaled $\boldsymbol{z}_K$, as $K \to \infty$. Here, the mixing distribution is assumed to converge to its limiting distribution slower than the multinomial distribution and thus the limiting distribution becomes that of the mixing distribution (Theorem 2).

Theorem 1 *We consider a sequence of J-vectors* $\boldsymbol{z}_K = (z_{K1}, \ldots, z_{KJ})'$. *Let* $\|\cdot\|$ *denote the Euclidean norm and assume that*

$$K\mathrm{E}(\|\boldsymbol{z}_K\|^2) = o(1)$$

when $K \to \infty$. Then

$$\frac{1}{\sqrt{K}}(\boldsymbol{x}_K - K\boldsymbol{\alpha}) \rightsquigarrow \mathcal{N}(\mathbf{0}, \mathrm{diag}(\boldsymbol{\alpha}) - \boldsymbol{\alpha}\boldsymbol{\alpha}'),$$

where $\rightsquigarrow$ denotes convergence in distribution.

Theorem 2 *Assume that*

$$\sqrt{\alpha_K}\boldsymbol{z}_K \rightsquigarrow \mathcal{D},$$

for some rate α_K and distribution $\mathcal{D}$, when $K \to \infty$. Assume further that $\alpha_K/K = o(1)$. Then

$$\frac{\sqrt{\alpha_K}}{K}(\boldsymbol{x}_K - K\boldsymbol{\alpha}) \rightsquigarrow \mathcal{D}.$$

The formal proofs are provided in the Appendix.

3 The special case of the Dirichlet-multinomial distribution

We apply Theorems 1 and 2 in the special case in which the mixing distribution is Dirichlet, i.e., when the class-specific probabilities of a multinomial model vary according to a Dirichlet distribution. We thus consider the following hierarchical model:

$$\boldsymbol{\pi}_K \sim \text{Dirichlet}(\alpha_{K1}, \ldots, \alpha_{KJ}),$$
$$\boldsymbol{x}_K|(K, \boldsymbol{\pi}_K) = (x_{K1}, \ldots, x_{KJ})|(K, \boldsymbol{\pi}_K) \sim \text{Multinomial}(K, \boldsymbol{\pi}_K),$$

where the second row indicates that the conditional distribution of counts $\boldsymbol{x}_K$ is multinomial of size K with a probability vector $\boldsymbol{\pi}_K$. We reparametrise the Dirichlet distribution in terms of $\tilde{\alpha}_j = \alpha_{Kj}/\alpha_K$, $j = 1, \ldots, J$, where $\alpha_K = \sum_{j=1}^{J} \alpha_{Kj}$, making the simplifying assumption that the $\tilde{\alpha}_j$'s are independent of K. Parameter α_K controls the heterogeneity of class probabilities as compared to the purely multinomial distribution. We call α_K sparsity, because the smaller its value, the more heterogeneous the class probabilities become between different samples, i.e., the more sparse the ensuing counts are. Note that we consider α_K as a sequence of values indexed by K and with the limit $\alpha_K \to \infty$ as $K \to \infty$. We assume that each $\tilde{\alpha}_j \alpha_K > 1$, and hence $\alpha_K > J$, so that the Dirichlet distribution has a mode $((\alpha_{K1}, \ldots, \alpha_{KJ}) - \mathbf{1}_J)/(\alpha_K - J)$ with respect to the Lebesgue measure. Here $\mathbf{1}_J$ is the J-vector $(1, \ldots, 1)'$. We denote the vector $(\tilde{\alpha}_1, \ldots, \tilde{\alpha}_J)'$ by $\tilde{\boldsymbol{\alpha}}$. We delineate two cases according to whether $K/\alpha_K = o(1)$, i.e., $K/\alpha_K \to 0$, or $\alpha_K/K = o(1)$, i.e., $\alpha_K/K \to 0$, when both $\alpha_K \to \infty$ and $K \to \infty$.

Corollary 1 *Let* $\boldsymbol{x}_K \sim \text{Multinomial}(K, \boldsymbol{\pi}_K)$ *and* $\boldsymbol{\pi}_K \sim \text{Dirichlet}(\alpha_K \tilde{\boldsymbol{\alpha}})$ *where* $\tilde{\boldsymbol{\alpha}} \in \mathcal{S}^{J-1}$. *If* $K/\alpha_K = o(1)$, *then*

$$\frac{1}{\sqrt{K}}(\boldsymbol{x}_K - K\tilde{\boldsymbol{\alpha}}) \rightsquigarrow \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\pi}}),$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\pi}} = \text{diag}(\tilde{\boldsymbol{\alpha}}) - \tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\alpha}}'$.

Corollary 1 follows as $K\|\boldsymbol{z}_K\|$ converges to zero (for details, see the Appendix).

Let $p_{Kj} = x_{Kj}/K$, $j = 1, \ldots, J$, denote the proportions (i.e. relative frequencies) based on the Dirichlet-multinomial counts $\boldsymbol{x}_K$. Denote $\mathbf{p} = (p_{K1}, \ldots, p_{KJ})'$. Under the conditions of Corollary 1, it follows immediately that

$$\sqrt{K}(\mathbf{p} - \tilde{\boldsymbol{\alpha}}) \rightsquigarrow \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\pi}}).$$

With a large enough total count K and α_K the distribution of the proportions can thus be viewed as normal and we indicate the normal approximation by writing

$$\mathbf{p} \approx \mathcal{N}\left(\tilde{\boldsymbol{\alpha}}, \frac{1}{K}\boldsymbol{\Sigma}_{\boldsymbol{\pi}}\right). \tag{1}$$

Corollary 2 *Under the scenario in Corollary 1, if* $\alpha_K/K = o(1)$, *then*

$$\frac{\sqrt{\alpha_K + 1}}{K}(\boldsymbol{x}_K - K\tilde{\boldsymbol{\alpha}}) \rightsquigarrow \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\pi}}).$$

Corollary 2 follows from the fact that under the assumption of unimodality, the sequence $\boldsymbol{\pi}_K$ has a limiting normal distribution (for details, see the Appendix). It follows from Corollary 2 that for large enough values of K and α_K we can write

$$\mathbf{p} \approx \mathcal{N}(\tilde{\boldsymbol{\alpha}}, \left(\frac{1}{\alpha_K + 1}\right) \boldsymbol{\Sigma}_{\boldsymbol{\pi}}). \tag{2}$$

The asymptotic normality of proportions holds when both $K \to \infty$ and $\alpha_K \to \infty$, regardless of which of the two parameters grows faster. As shown above, when $\alpha_K \gg K$, the asymptotic behaviour of the distribution of the proportions is dominated by the asymptotic normality of the multinomial distribution. By contrast, when $K \gg \alpha_K$, the behaviour is dominated by the asymptotic normality of the mixing distribution (Dirichlet). Interestingly, the limiting distributions in Corollary 1 and Corollary 2 are the same. The difference between the two scenarios is only in the rate of convergence.

The results (1) and (2) can be combined by writing the approximation as

$$\mathbf{p} \approx \mathcal{N}(\tilde{\boldsymbol{\alpha}}, \boldsymbol{\Sigma}_{\boldsymbol{p}}), \tag{3}$$

where

$$\boldsymbol{\Sigma}_{\boldsymbol{p}} = \frac{1}{K}\left(\frac{\alpha_K + K}{\alpha_K + 1}\right) \boldsymbol{\Sigma}_{\boldsymbol{\pi}}. \tag{4}$$

In particular, $\boldsymbol{\Sigma}_{\boldsymbol{p}}$ reduces to those in expressions (1) or (2) when $\alpha_K \gg K$ or $K \gg \alpha_K$, respectively. The formulation above is motivated by the fact that $\boldsymbol{\Sigma}_{\boldsymbol{p}}$ is the variance-covariance matrix of relative frequencies of Dirichlet-multinomial counts, as found easily by the law of total variance. Of note, the excess variability of proportions $\mathbf{p}$ under the Dirichlet-multinomial distribution as compared to those under the multinomial distribution is

$$(\alpha_K + K)/(\alpha_K + 1). \tag{5}$$

We conclude the section with two technical remarks. First, Eq. (1) and Corollaries 1, 2 might at first seem counterintuitive in their claims that random points on the simplex (which is a compact set) should have limiting normal distributions (whose support is not a compact set). However, closer inspection reveals that the result is entirely natural. Firstly, in Eq. (1), the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\pi}}$ satisfies $\boldsymbol{\Sigma}_{\boldsymbol{\pi}} \mathbf{1}_J = \mathbf{0}_J$. This means that the limiting distribution has zero variation in the direction orthogonal to the simplex surface, and, accordingly, the only way the approximation can "escape" the simplex is in directions parallel to the simplex surface. Moreover, the variance of the limiting distribution decreases when K grows, meaning that in the limit the distribution has no mass outside of the simplex (even along directions parallel to the simplex), and a limiting normal distribution is achieved with the appropriate scaling. We have illustrated this behaviour in Fig. 1, which shows how the equidensity contours of the approximate normal distribution shrink on the simplex surface

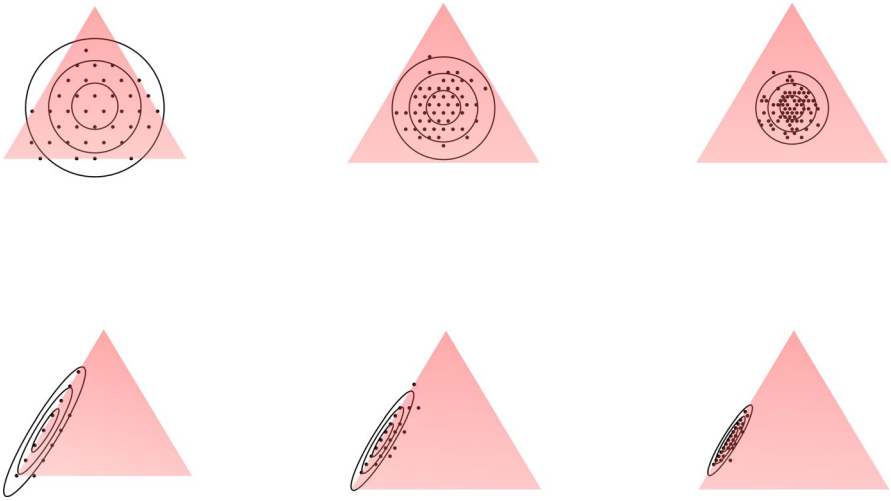

Fig. 1 The points depict samples of 100 random Dirichlet-multinomial realisations of $\boldsymbol{x}_K/K$ when either $\tilde{\boldsymbol{\alpha}} = (1/3, 1/3, 1/3)$(top row) or $\tilde{\boldsymbol{\alpha}} = (0.01, 0.30, 0.69)$(bottom row), and $\alpha_K = K^{1.5}$ and K is either 10 (left), 20 (middle) or 40 (right). The ellipses correspond to Mahalanobis distances of one, two and three units from the mean and are based on the limiting normal distributions predicted by Corollary 1 and Eq. (1)

in two different settings as K is increased, eventually (in the limit) leading to all probability mass being contained on the simplex. In the first setting, where the distribution is symmetrically located around the centre of the simplex, the majority of the probability mass of the distribution is quickly concentrated within the simplex when K grows. By contrast, in the second setting where the distribution is located close to the border of the simplex, much larger total counts K are required. These plots hence serve to illustrate that the quality of the asymptotic approximations depends not only on K and α_K but also on the structure of the data at hand.

Secondly, keeping in mind that our asymptotic results pertain to individual observations $\boldsymbol{x}_K$, one would optimally want to have the simultaneous convergence of all elements of a sample $(\boldsymbol{x}_{K,1}, \ldots, \boldsymbol{x}_{K,n})$ to limiting normal distributions, where $\boldsymbol{x}_{k,i}$ are i.i.d. realisations of $\boldsymbol{x}_K$ and both $K, n \to \infty$. However, such double-asymptotic results are beyond our scope, and in practice, it is sufficient that the previous holds for any fixed n (with $K \to \infty$), a fact that follows instantly from our results by the independence of the $\boldsymbol{x}_{K,i}$.

3.1 Overdispersed total count

Next, we extend the Dirichlet-multinomial model to allow for overdispersion not only in the proportions across the samples but also in the total count. In general, empirical data on read counts often exhibit both types of overdispersion. A typical model for the total count is the negative binomial distribution (i.e. overdispersed Poisson). However, we here formulate the model of the total count using the log-normal distribution for it will allow an explicit expression of the moments of the marginal distribution

of the proportions. While the log-normal distribution is continuous, the values of the total counts can in practice be rounded to the nearest integer.

We denote the distribution of K as lognormal(μ,σ^2), with the expectation and variance

$$\begin{aligned} &\mathrm{E}(K) = \exp(\mu + \sigma^2/2) \\ &\mathrm{Var}(K) = (\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2). \end{aligned} \tag{6}$$

It follows from (3) and (6) that the marginal expectation of the vector of proportions $\mathbf{p}$ is

$$\mathrm{E}(\mathbf{p}) = \mathrm{E}_K\left(\mathrm{E}\left(\mathbf{p}|K\right)\right) = \mathrm{E}_K(\tilde{\boldsymbol{\alpha}}) = \tilde{\boldsymbol{\alpha}}.$$

Similarly, by the law of total covariance, the variance-covariance matrix of $\mathbf{p}$ is

$$\begin{aligned} \boldsymbol{\Sigma}_{\boldsymbol{p}} &= \mathrm{E}_K\left(\mathrm{Cov}(\mathbf{p}, \mathbf{p}'|K)\right) + \mathrm{Cov}_K\left(\mathrm{E}\left(\mathbf{p}|K\right), \mathrm{E}\left(\mathbf{p}'|K\right)\right) \\ &= \mathrm{E}_K\left(\frac{1}{K}\left(\frac{\alpha_K + K}{\alpha_K + 1}\right)\boldsymbol{\Sigma}_{\boldsymbol{\pi}}\right) + \mathrm{Cov}_K\left(\tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\alpha}}'\right) \\ &= \boldsymbol{\Sigma}_{\boldsymbol{\pi}}\left(\frac{1}{\alpha_K + 1}\right)\left(\alpha_K \exp(-\mu + \sigma^2/2) + 1\right) = \boldsymbol{\Sigma}_{\boldsymbol{\pi}}\gamma(\alpha_K, \mu, \sigma^2), \end{aligned} \tag{7}$$

where we denote

$$\gamma(\alpha_K, \mu, \sigma^2) = \left(\frac{1}{\alpha_K + 1}\right)\left(\alpha_K \exp(-\mu + \sigma^2/2) + 1\right). \tag{8}$$

Proposition 3 *When the distribution of the total count is lognormal(μ, σ^2), a normal approximation for the distribution of the proportions $\mathbf{p}$ is given by*

$$\mathbf{p} \approx \mathcal{N}(\tilde{\boldsymbol{\alpha}}, \gamma(\alpha_K, \mu, \sigma^2)\boldsymbol{\Sigma}_{\boldsymbol{\pi}}).$$

We here have used the property of the lognormal distribution according to which the distribution of $1/K$ is lognormal($-\mu$, σ^2) if the distribution of K is lognormal(μ, σ^2). The excess variability of proportions under lognormal total count, in comparison to the multinomial distribution with total count K chosen to correspond to the median of the lognormal distribution, is $\exp(\mu)\gamma(\alpha_K, \mu, \sigma^2)$. In the limit $\alpha_K \to \infty$, this ratio is $\exp(\mu)\exp(-\mu + \sigma^2/2) = \exp(\sigma^2/2)$. We investigate the performance of the proposed approximation in Sect. 5.4.

4 Ilr coordinates and their asymptotic normality

4.1 Contrasts and the ilr coordinates

Proportions based on counts are here considered as a composition, constrained by the unit-sum condition $\sum_{j=1}^{J} p_j = 1$. To transform the J-part proportions to $(J-1)$ -dimensional Euclidean coordinates, we apply the isometric log-ratio (ilr) transformation (Egozcue et al. 2003).

The first step is to define an appropriate set of contrasts between the J components. There is no canonical basis for choosing the contrasts (Bacon-Shone 2011). A popular approach for this is the sequential binary partition approach (SBP), where the parts of the composition are contrasted against each other in a hierarchical manner, resulting in coordinates called *balances* (Pawlowsky-Glahn et al. 2015). The partitions are encoded using a $J \times (J-1)$ sign matrix. The chosen partitions can be based on, for example, known relationships between the parts of the composition so that the resulting coordinates can be interpreted in a straightforward manner.

One way to define the partition is using pivot coordinates, where one part is contrasted against all other classes, and the other parts of the composition are sequentially contrasted against the remaining parts (Filzmoser et al. 2018). This leads to the following $J \times (J-1)$ SBP matrix:

$$\boldsymbol{\Psi} = (\psi_{jk}) = \begin{bmatrix} +1 & 0 & 0 & \dots & 0 \\ -1 & +1 & 0 & \dots & 0 \\ -1 & -1 & +1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & -1 & \vdots & +1 \\ -1 & -1 & -1 & \vdots & -1 \end{bmatrix}. \tag{9}$$

We here refer to the matrix used to build pivot coordinates as pivotal.

Let $\mathbf{m}$ denote the vector of ilr coordinates. With any given SBP matrix $\boldsymbol{\Psi}$, the ilr transformation is obtained as follows (Graffelman et al. 2015):

$$\mathbf{m} = \text{ilr}(\mathbf{p}) = \mathbf{V}'\ln(\mathbf{p}),$$

where the contrast matrix $\mathbf{V}$ is based on $\boldsymbol{\Psi}$:

$$(\mathbf{V})_{jk} = \psi_{jk}\sqrt{\frac{n_k^+ n_k^-}{n_k^+ + n_k^-}}\left[\frac{1}{n_k^+}\right]^{\text{I}[\text{sign}(\psi_{jk})=+]}\left[\frac{1}{n_k^-}\right]^{\text{I}[\text{sign}(\psi_{jk})=-]}.$$

Here, n_k^+ and n_k^- are the numbers of cells with values $+1$ and -1 in the kth column of $\boldsymbol{\Psi}$. In general, the above definition means that ilr coordinates can be calculated as

$$m_k = \sqrt{\frac{n_k^+ n_k^-}{n_k^+ + n_k^-}} \ln \frac{\mathrm{gm}(p_k^+)}{\mathrm{gm}(p_k^-)}, k = 1, \ldots, J-1,$$

where $\mathrm{gm}(p_k^+)$ denotes the geometric mean of the proportions in the classes denoted by $+1$. Whereas the original proportions lie in the simplex, the transformed coordinates reside in the Euclidean space.

4.2 Asymptotic distribution of the ilr coordinates and their normal approximation

The asymptotic normality of the ilr coordinates when based on the (ilr) transformation of purely multinomial counts has been shown earlier (Graffelman et al. 2015). Based on our Theorems 1 and 2, it follows immediately by the delta method that the asymptotic normality of the ilr coordinates holds under more general conditions, which allow extra-multinomial variation. To demonstrate this, we here derive an explicit asymptotic normal approximation to the ilr coordinates under the special case of Dirichlet-multinomial counts.

Based on the delta method, a transformation g of proportions $\mathbf{p}$ based on the Dirichlet-multinomial mixture has the following asymptotic approximation:

$$g(\mathbf{p}) \approx \mathcal{N}\left(g(\mathrm{E}(\mathbf{p})), \left(\frac{\partial g(\mathbf{p})}{\partial \mathbf{p}}\right) \boldsymbol{\Sigma}_{\boldsymbol{p}} \left(\frac{\partial g(\mathbf{p})}{\partial \mathbf{p}}\right)'\right),$$

where $\boldsymbol{\Sigma}_{\mathbf{p}}$ is the variance-covariance matrix (4) and the derivatives are evaluated at $\mathrm{E}(\mathbf{p}) = \tilde{\boldsymbol{\alpha}}$. For $g(\mathbf{p}) = \mathrm{ilr}(\mathbf{p})$, the derivatives are

$$\frac{\partial \mathrm{ilr}(\mathbf{p})}{\partial \mathbf{p}} = \frac{\partial \mathbf{V}' \ln(\mathbf{p})}{\partial \mathbf{p}} = \mathbf{V}' \mathbf{D}_{\tilde{\boldsymbol{\alpha}}}^{-1}.$$

Based on the delta method, we thus arrive at the asymptotic distribution for the ilr coordinates.

Corollary 3 *The asymptotic normal approximation of the ilr coordinates is:*

$$\mathrm{ilr}(\mathbf{p}) \approx \mathcal{N}\left(\mathrm{ilr}(\tilde{\boldsymbol{\alpha}}), \mathbf{V}' \mathbf{D}_{\tilde{\boldsymbol{\alpha}}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{p}} \mathbf{D}_{\tilde{\boldsymbol{\alpha}}}^{-1} \mathbf{V}\right). \tag{10}$$

Under multinomial counts, the variance-covariance matrix is simplified to $(1/K)\mathbf{V}'\mathbf{D}_{\tilde{\boldsymbol{\alpha}}}^{-1}\mathbf{V}$(Graffelman et al. 2015). In case of Dirichlet-multinomial counts, we can similarly write the variance-covariance matrix as $(1/K)(\alpha_K + K)/(\alpha_K + 1)\mathbf{V}'\mathbf{D}_{\tilde{\boldsymbol{\alpha}}}^{-1}\mathbf{V}$.

In practice, $\mathrm{ilr}(\tilde{\boldsymbol{\alpha}})$ may deviate considerably from $\mathrm{E}(\mathrm{ilr}(\mathbf{p}))$. For any positive scalar random variable X, the first-order Taylor approximation of the expectation of $\ln(X)$ is

$$\mathrm{E}(\ln(X)) \simeq \ln(\mathrm{E}(X)) - \frac{\mathrm{Var}(X)}{2\mathrm{E}(X)^2},$$

which holds well for large X. Applying the above approximation separately for each element of $\mathbf{p}$, we obtain

$$\mathrm{E}(\ln(\mathbf{p})) \simeq \ln(\tilde{\boldsymbol{\alpha}}) - (1/2)(1/K)((\alpha_K + K)/(K\alpha_K + K))\ \boldsymbol{\lambda} \circ \boldsymbol{\beta},$$

where $\boldsymbol{\beta} = (1/\tilde{\alpha}_1^2, \ldots, 1/\tilde{\alpha}_J^2)$, $\boldsymbol{\lambda}$ is a vector of the diagonal elements of $\boldsymbol{\Sigma}_{\boldsymbol{\pi}}$, and $\circ$ marks element-wise multiplication. Using this alternative expression, the asymptotic approximation of ilr coordinates based on Dirichlet-multinomial counts is

$$\begin{aligned} &\mathrm{ilr}(\mathbf{p}) \approx \mathcal{N}(\boldsymbol{\eta}, \boldsymbol{\Xi}), \text{ where} \\ &\boldsymbol{\eta} = \mathbf{V}'(\ln(\tilde{\boldsymbol{\alpha}}) - (1/2)(1/K)((\alpha_K + K)/(K\alpha_K + K))\ \boldsymbol{\lambda} \circ \boldsymbol{\beta}) \text{ and} \\ &\boldsymbol{\Xi} = (1/K)(\alpha_K + K)/(\alpha_K + 1)\mathbf{V}'\mathbf{D}_{\tilde{\boldsymbol{\alpha}}}^{-1}\mathbf{V}. \end{aligned} \tag{11}$$

Finally, if the total count follows the log-normal distribution instead of being fixed, the variance-covariance matrix $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{p}}$ of the proportions is given by (7) and thus the covariance matrix in (11) depends on parameters μ, α_K and σ^2. Thus, the asymptotic approximation of the ilr coordinates is now

$$\begin{aligned} &\mathrm{ilr}(\mathbf{p}) \approx \mathcal{N}\left(\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\Xi}}\right), \text{ where} \\ &\tilde{\boldsymbol{\eta}} = \mathbf{V}'(\ln(\tilde{\boldsymbol{\alpha}}) - (1/2)((\alpha_K \exp(-\mu + \sigma^2/2) + 1)/(\alpha_K + 1))\ \boldsymbol{\lambda} \circ \boldsymbol{\beta}) \text{ and} \\ &\tilde{\boldsymbol{\Xi}} = \gamma(\alpha_K, \mu, \sigma^2)\mathbf{V}'\mathbf{D}_{\tilde{\boldsymbol{\alpha}}}^{-1}\mathbf{V}. \end{aligned} \tag{12}$$

5 Simulation study

In this section, we investigate the applicability of the normal approximation (11) of the ilr coordinates under different levels of sparsity α_S and the total count K. In the following, the sparsity parameter is denoted as α_S instead of α_K as here it will not represent any formal link to the total count K. We also investigate how different levels of variability of the total count influence the performance of the approximation.

5.1 Simulation setup

We consider four different data generating distributions (Table 1): (a) multinomial with a fixed total count and fixed class probabilities, (b) Dirichlet-multinomial with a fixed total count and variable class probabilities, (c) lognormal-multinomial with a variable total count and fixed class probabilities and (d) lognormal-Dirichlet-multinomial with variable total count and variable class probabilities. Table 2 presents the values of the model parameters as used in the simulation scenarios. The number of classes $J = 5$ and the class-specific expected probabilities $\tilde{\boldsymbol{\alpha}}' = (0.01, 0.04, 0.15, 0.3, 0.5)$ were fixed. The SBP matrix was built in a pivotal manner as in Eq. (9), which will

Table 1 Four data generating distributions and their parameters

		Total count K	
		Fixed	Lognormal
Prob $\boldsymbol{\pi}$	Fixed	a) Multinomial(K, $\boldsymbol{\pi}$)	c) Lognormal-multinomial(μ, σ^2, $\boldsymbol{\pi}$)
	Dirichlet	b) Dir-multinomial(K, $\boldsymbol{\alpha}$)	d) Lognormal-Dir-multinomial(μ, σ^2, $\boldsymbol{\alpha}$)

Table 2 Parameters of the simulation study

Parameter	Values	Note
J	5	Number of classes
$\boldsymbol{\pi}$; $\tilde{\boldsymbol{\alpha}}$	(0.01, 0.04, 0.15, 0.3, 0.5)	Class-specific expected probabilities
α_S	101; 1000; 10000; 100000; 1000000	Sparsity parameter (b,d)
K; $\exp(\mu)$	101; 1000; 10000; 100000; 1000000	Total count (a,b); median total count (c,d)
σ^2	0.1; 1.0	Variability of logarithm of total count (c,d)

The sparsity parameter α_S only applies when $\boldsymbol{\pi} \sim$ Dirichlet. Parameter σ^2 only applies when $K \sim$ lognormal. The total count is either fixed at K or has a lognormal distribution with median $\mu = \ln(K)$. The four different data generating distributions are indicated by a–d (see Table 1)

allow investigation of coordinates involving both rare and more abundant classes, as well as only abundant classes. We varied the value of α_S, which defines the sparseness of the count data. The values for α_S were chosen so that the unimodality condition was fulfilled, i.e., each $\tilde{\alpha}_j \alpha_S > 1$. We also varied the total count K or, under lognormal counts, its median $\exp(\mu)$ and variability σ^2. The coefficient of variation (CV) of the lognormally distributed total count is $\sqrt{\exp(\sigma^2) - 1}$, independent of μ. The values 0.1 and 1.0 for σ^2 correspond to CV's of 0.32 and 1.31, respectively. In empirical microbiome data, the CV typically falls within this range (Aatsinki et al. 2018; Keskitalo et al. 2021; Aatsinki et al. 2019, 2020).

Table 3 presents the excess variability of the class-specific proportions and, subsequently, of the ilr coordinates under each data generating mechanism and each parameter setting. The excess variability was calculated as the ratio of the elements of the variance-covariance matrix of the proportions to those under the multinomial distribution, based on equations (5) and (8).

Each simulation consisted of 10000 independent draws of counts from their data generating distribution. Based on the 10000 draws, we obtained the simulated expectations and variance-covariance matrices of the ilr coordinates as approximations to their true values under finite K and finite α_S.

It is obvious that with a finite total count K, zero-count observations may occur and $\ln(p_i)$, when based on such an observation, is undefined and the ilr coordinate cannot be calculated. In general, zero counts can be essential zeros, which in the context of microbial data means that the zero-count taxa are absent from the sample, or rounded zeros, corresponding to an observed proportion that is smaller than e.g.

Table 3 Excess variability in class-specific proportions (p_j) under some selected simulation scenarios with varying values of sparsity parameter α_S and total count K or, in case of lognormally distributed total count, its median $\exp(\mu)$ and variability σ^2

DGD	Eq	α_S	σ^2	$K=101$	$K=1000$	$K=10000$	$K=100000$	$K=1000000$
(a) Mn	–	–	–	1.00	1.00	1.00	1.00	1.00
(b) Dir-Mn	(5)	101	–	1.98	10.79	99.03	981.38	9804.91
(b) Dir-Mn	(5)	1000	–	1.10	2.00	10.99	100.90	1000.00
(b) Dir-Mn	(5)	10000	–	1.01	1.10	2.00	11.00	100.99
(b) Dir-Mn	(5)	100000	–	1.00	1.01	1.10	2.00	11.00
(b) Dir-Mn	(5)	1000000	–	1.00	1.00	1.01	1.10	2.00
(c) LN-Mn	(8)	–	0.1	1.05	1.05	1.05	1.05	1.05
(d) LN-Dir-Mn	(8)	101	0.1	2.03	10.84	99.08	981.43	9804.96
(d) LN-Dir-Mn	(8)	1000	0.1	1.15	2.05	11.04	100.95	1000.05
(d) LN-Dir-Mn	(8)	10000	0.1	1.06	1.15	2.05	11.05	101.04
(d) LN-Dir-Mn	(8)	100000	0.1	1.05	1.06	1.15	2.05	11.05
(d) LN-Dir-Mn	(8)	1000000	0.1	1.05	1.05	1.06	1.15	2.05
(c) LN-Mn	(8)	–	1	1.65	1.65	1.65	1.65	1.65
(d) LN-Dir-Mn	(8)	101	1	2.62	11.44	99.67	982.02	9805.55
(d) LN-Dir-Mn	(8)	1000	1	1.75	2.65	11.64	101.55	1000.65
(d) LN-Dir-Mn	(8)	10000	1	1.66	1.75	2.65	11.65	101.64
(d) LN-Dir-Mn	(8)	100000	1	1.65	1.66	1.75	2.65	11.65
(d) LN-Dir-Mn	(8)	1000000	1	1.65	1.65	1.66	1.75	2.65

The variance-covariance matrix elements are compared against those of multinomial proportions, based on equations (5) or (8). *DGD* Data generating distribution (see Table 1), *Mn* Multinomial, *Dir* Dirichlet, *LN* Lognormal

a detection limit. The third option is called count zeroes, meaning that the taxa may be present in the microbial community but not observed in the sample due to finite or insufficiently large sample (Martín-Fernández et al. 2014, 2003). Our simulation model is basically developed on the notion that any zeroes are count zeroes. A common method in microbiome studies is to replace any zero counts with pseudo-counts, often set to 0.5 (Zhang et al. 2021; Sohn and Li 2019). This approach was used in our main simulation settings in case of zero-count observations. However, as fixed pseudo-counts have the potential to distort the structure of the data and the ratios (Fry et al. 2000), alternative methods for zero replacement have been suggested (Martín-Fernández et al. 2014). As an additional analysis, we applied a method called *Bayesian multiplicative zero replacement* using the function `cmultRepl` in R package zCompositions (Palarea-Albaladejo and Martín-Fernández 2015). We used this method to impute all zero counts in a Bayesian manner by replacing them by their posterior expectations under a Dirichlet prior. The results of this additional analysis are presented in Sect. 5.4.2.

We assessed the normality of the coordinates by comparing the quantiles of the simulated ilr coordinates to the theoretical quantiles using normal Q–Q-plots and investigated the performance of approximation (11) by comparing its expectation and eigenvalues of the variance-covariance matrix to their finite-K/finite-α_S counterparts as obtained by simulation. For each combination of α_S and K, the comparison was made by plotting the logarithmic ratios of the simulated and asymptotic expectations and eigenvalues.

5.2 Illustration of the composition of counts and proportions

Before considering the distribution of the ilr coordinates, we illustrate the composition of counts under selected values of K and α_S(Fig. 2). As the total count K is here fixed, the same compositions apply to class proportions. We investigate scenarios with either small ($K = 101$) or large ($K = 1000000$) total count, and with either high sparsity ($\alpha_S = 101$), intermediate sparsity ($\alpha_S = 10000$), or no sparsity (multinomial). With a large total count, the counts exhibit less variation than under a small total count. With increasing values of α_S, the distribution of counts approaches that of multinomial and the compositions exhibit decreasing amounts of heterogeneity. Following the choice of the α_S parameters, the distributions of the counts are unimodal. In particular, for each combination of α_S and K, the most common proportion is p_5 in nearly all samples shown in Fig. 2, and the composition of the proportions under $\alpha_S = 10000$ does not largely differ from that under the multinomial distribution. An alternative presentation of the proportions is presented in Figure S1, where the indivual observations are sorted in descending order of the three rarest classes.

5.3 Normality of the ilr coordinates and the performance of the normal approximation

To understand under which scenarios the asymptotic normality of the ilr coordinates holds, we explored the normality of the first and the fourth ilr coordinates under selected parameter combinations using Q–Q plots (Fig. 3). The last ilr coordinate

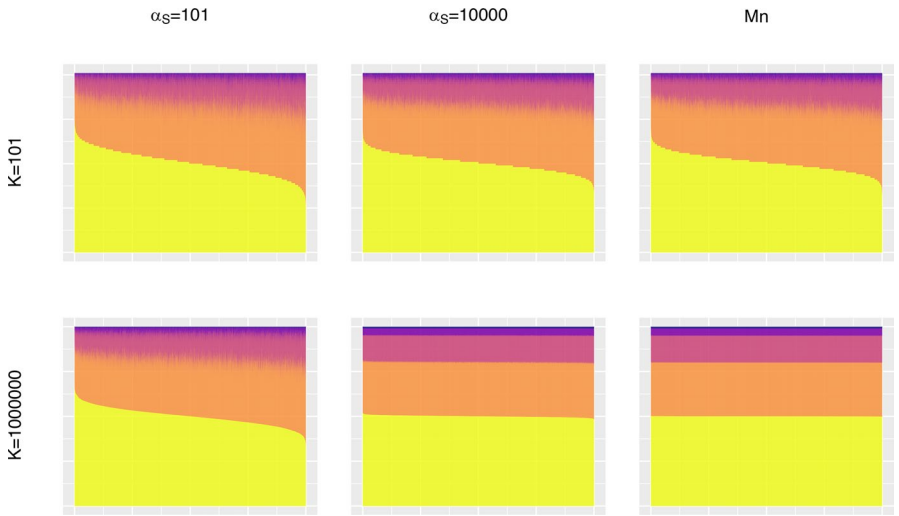

Fig. 2 Compositions of counts under the multinomial (data generating distribution **a**; right-hand column) and Dirichlet-multinomial (data generating distribution **b**; left-hand and middle column) distributions when the total count K is either 101 (upper panels) or 1000000 (lower panels). Each vertical line represents one observation with the five counts stacked from K_1(upmost) to K_5(lowest) and indicated by the colours corresponding to the five class-specific counts with class probabilities $\tilde{\boldsymbol{\alpha}} = (0.01, 0.04, 0.15, 0.3, 0.5)$. The observations are sorted in the descending order of K_5. As the total count is fixed, the same compositions apply to proportions

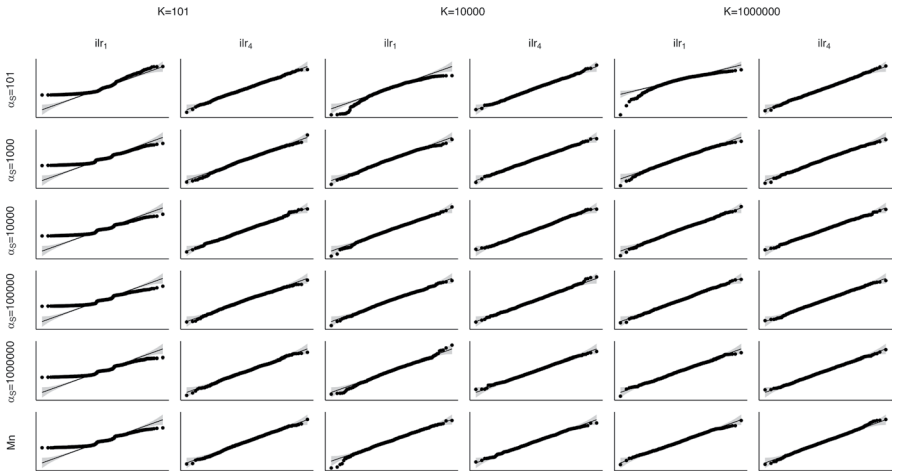

Fig. 3 Quantile-quantile plot for the first (ilr_1) and fourth (ilr_4) ilr coordinates under different parameter combinations under multinomial and Dirichlet-multinomial counts. The horizontal axis corresponds to theoretical and the vertical axis to simulated sample quantiles

(ilr_4), consisting only of the two most common classes, tends to be approximately normal with all presented combinations of K and α_S. The first coordinate, contrasting the rarest class against all other classes, is normally distributed when K is large and there is very little or no sparsity. However, with small α_S or small K, the assumption of normality does not hold as well. This stems from the distribution of the proportions of the first (least abundant) class, as presented in Fig. S2.

We next examined the performance of approximation (11) through comparisons with the simulated distribution of the ilr coordinates under finite K and α_S. Under multinomial counts, the asymptotic approximation of the expected values works well when the total count K is larger than 100 (Fig. 4, Mn), corresponding to the findings in Graffelman et al. (Graffelman et al. 2015). Also under the Dirichlet-multinomial distribution, the approximation holds quite well when both $\alpha_S \to \infty($ i.e., when the distribution of counts approaches multinomial) and $K \to \infty($ Fig. 4, connected points). In this case, the bias remains small and is mostly $< 1\%$. When either α_S or K is small ($\alpha_S = 101$ or $K = 101$), the asymptotic expectations, especially for the first coordinate, deviate more from the simulated value with the absolute bias up to 30 $\%$. Interestingly, the asymptotic approximation appears to be more accurate for the last three ilr coordinates than the first one. The accuracy depends on which proportions are contrasted in each of the coordinates. While the first coordinate involves also the rarest classes, the last three coordinates are only based on the more abundant ones, reflecting that the Taylor approximation utilised here performs well for larger abundances.

For each eigenvalue of the variance-covariance matrix, the values based on approximation (11) correspond to the simulated values when both α_S and K are large (Fig. 5). When either α_S or K is small, however, the approximation does not work as well. In general, the first eigenvalue of the asymptotic variance-covariance matrix differs the most from the simulated under finite K and α_S. Especially with the small-

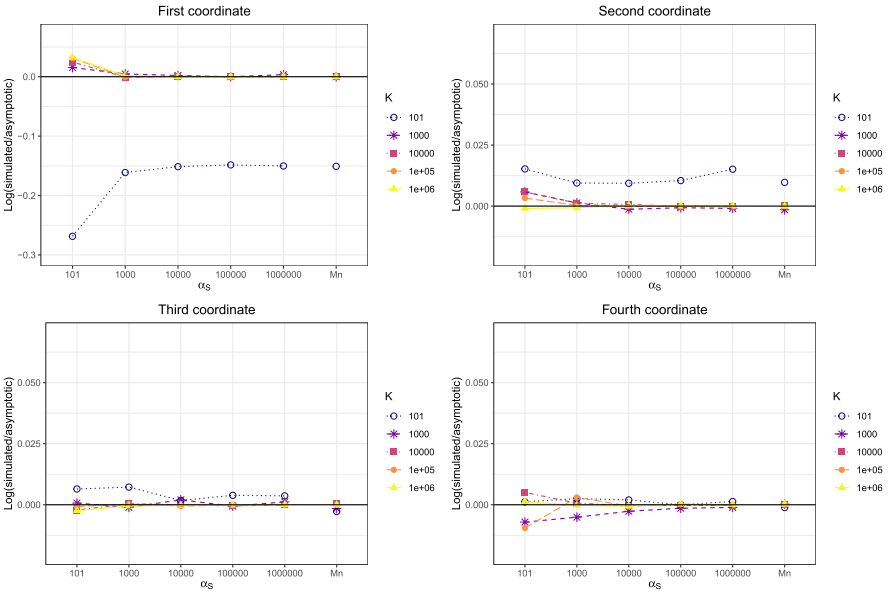

Fig. 4 Log-ratios of the expectations under finite K and α_S, based on Monte Carlo simulation, and the asymptotic expected values (Eq. 11) for the four ilr coordinates under multinomial (Mn; data generating distribution **a** in Table 1) and Dirichlet-multinomial counts (connected dots; data generating distribution **b** in Table 1). Value 0 means perfect correspondence between the simulated and asymptotic values. Note that the scale of the vertical axis for the first coordinate is different from the rest

est K, the asymptotic first eigenvalue deviates from the simulated value, even under multinomial counts.

To summarise the above findings, we note that although the excess variability with respect to the multinomial distribution in the class-specific proportions becomes excessively large with increasing total count K (Table 3), the asymptotic approximations for both the expected values and eigenvalues perform well. By contrast, the approximation is not as good with small values of K, even when the excess variability is relatively small. It can thus be concluded than when K is large, the distribution of the proportions can diverge from the multinomial distribution without compromising the near-normality of the ilr coordinates. On the other hand, when K is small, a much smaller departure from the multinomial distribution worsens the performance of the approximation. It is of note that the results regarding the normality presented in the Q–Q-plots and the results regarding the performance of the approximation are aligned. Under those scenarios where the distributions of the least abundant proportions and hence the ilr coordinates deviates from the normal, the approximation also tends to have poor performance.

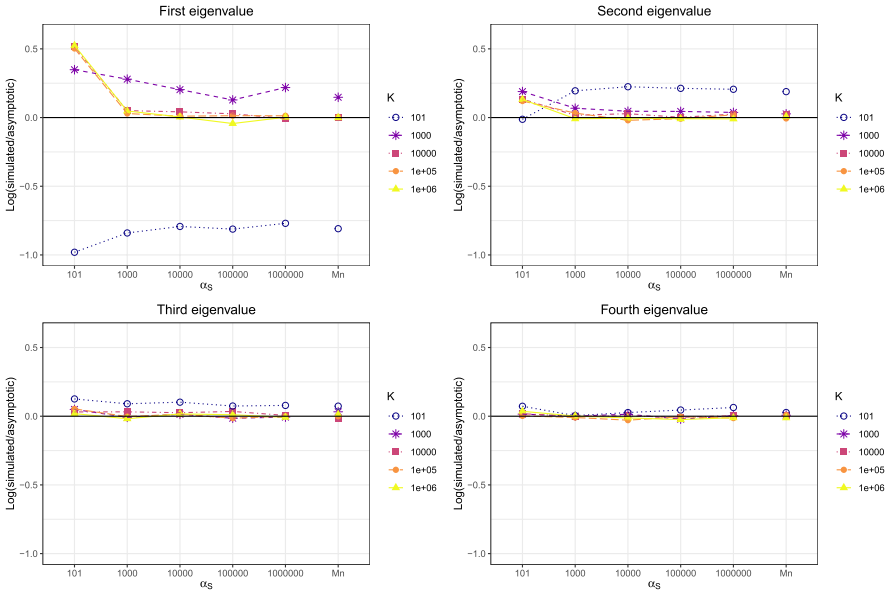

Fig. 5 Log-ratios of the eigenvalues under finite K and α_S, based on Monte Carlo simulation, and the eigenvalues of the asymptotic variance-covariance matrix (Eq. 11) for the four ilr coordinates under multinomial (Mn; data generating distribution **a** in Table 1) and Dirichlet-multinomial counts (connected dots; data generating distribution **b** in Table 1). Value 0 means perfect correspondence between the simulated and asymptotic values

5.4 Performance of the approximation in further settings

5.4.1 Normal approximation based on multinomial counts

We next demonstrate the pitfalls of using a normal approximation that is based on assuming purely multinomial counts when the underlying counts actually exhibit extra-multinomial variation. Letting $\alpha_S \to \infty$ in (11), the approximation becomes (see also (Graffelman et al. 2015))

$$\text{ilr}(\mathbf{p}) \approx \mathcal{N}(\mathbf{V}'\ln(\mathbf{p}), \frac{1}{K}\mathbf{V}'\mathbf{D}_{\tilde{\alpha}}^{-1}\mathbf{V}). \tag{13}$$

We here present results only for the Dirichlet-multinomial data generating distribution. However, conclusions were similar when the total count was lognormally distributed instead of being fixed (data not shown).

Figure 6 compares the simulated (i.e. finite-K and finite-α_S) expectations of the ilr coordinates to those of the multinomial approximation (13) in the same simulation scenarios as in Fig. 4. When the counts are sparse (small α_S), the approximation does not hold even under when the total count K is large (the bias between simulated and asymptotic expected values is up to 20 %). As expected, when the distribution of the counts approaches multinomial, approximation (13) improves.

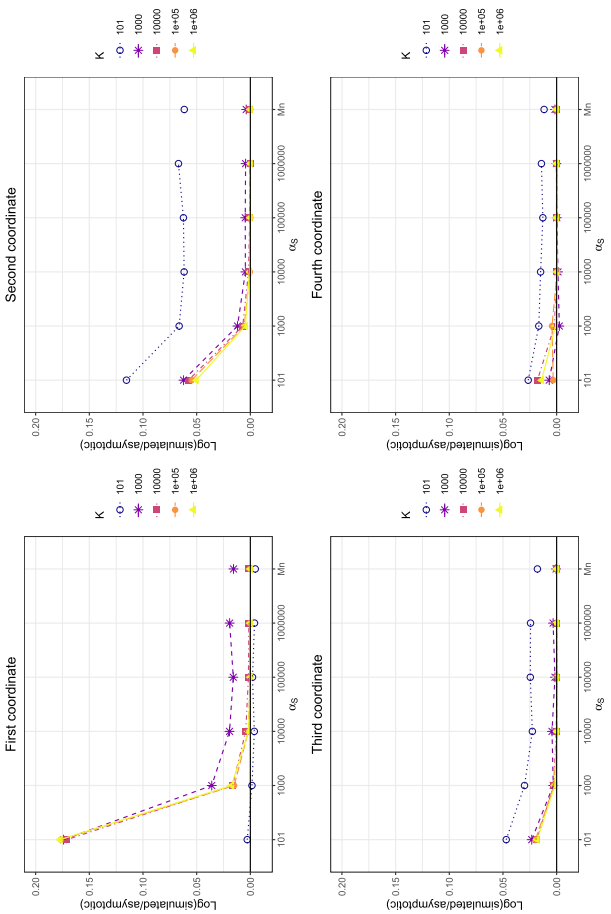

Fig. 6 Log-ratios of the expectations under finite K and α_S, based on Monte Carlo simulation, and normal approximation based on multinomial counts under multinomial (Mn) and Dirichlet-multinomial (connected dots) counts. Value 0 means perfect correspondence between the simulated and asymptotic values. Note that the scale differs from that in Fig. 4

Of note, the asymptotic approximation of the expected value used here is not based on a Taylor approximation. Comparing the results presented in Figs. 4 and 6 thus provides evidence on the additional value of using the Taylor approximation for the expected value.

Figure 7 compares the simulated eigenvalues of the variance-covariance matrix of the ilr coordinates with the corresponding eigenvalues under approximation (13) in the same simulation scenarios as in Fig. 5. The asymptotic values deviate from the simulated ones when the counts are sparse (small α_S). However, even under the largest value of α_S, the approximation is not particularly good. Somewhat unintuitively, in contrast to the expected values, the performance of the approximation decreases with increasing K. This stems from the excess variability in the Dirichlet-multinomial distribution, characterised in equation (5). When K increases, the excess variability also increases.

5.4.2 Bayesian replacement of zero-count observations

We next present an additional simulation study where all zero-count observations in the simulated data were imputed using Bayesian multiplicative zero replacement instead of a fixed pseudo-count of 0.5 (Palarea-Albaladejo and Martín-Fernández 2015). Under large K and α_S(i.e. less sparsity), there were none or only a few zero-count observations in the simulated data. Furthermore, there were no zero-count observations in the two most abundant classes under any of the scenarios. With small K, especially the least abundant class included zero-count observations, their amount increasing when the sparsity increases (i.e. α_S decreases) (Table S1). Our interpretations below apply on the simulation scenarios with non-negligible amount of zero-count observations.

When zeroes were imputed with Bayesian replacement instead of a fixed pseudo-count of 0.5, approximation (11) was in closer agreement with the estimated expected values of the first ilr coordinate under all values for α_S when $K = 101$(cf. Figures S4 and 4 for Bayesian replacement and the fixed pseudo-count, respectively). It is also of note that the direction of the bias differs between the two alternative methods. However, when $K = 1000$ and the data are sparse ($\alpha_S = 101$), the fixed pseudo-count yields slightly better correspondence with the approximation for the first coordinate. The fixed pseudo-count also performs slightly better for the second ilr coordinate. The simulated expected values obtained for the third and fourth ilr coordinates are identical between the two approaches due to lack of zero-count observations (Table S1).

For the first eigenvalue of the variance-covariance matrix of the ilr coordinates, Bayesian replacement aligns more closely with the approximation (11), the direction of bias being different to that of the fixed pseudo-count approach (cf. Figure S5 for Bayesian replacement and Fig. 5 for the fixed pseudo-count). The fixed pseudo-count has slightly better alignment with the approximation of the second eigenvalue. The differences between the two approaches are negligible between the third and fourth eigenvalues.

While the Bayesian replacement performs better in terms of the moments, interestingly, with the smallest total count ($K = 101$), the distribution of the first coordi-

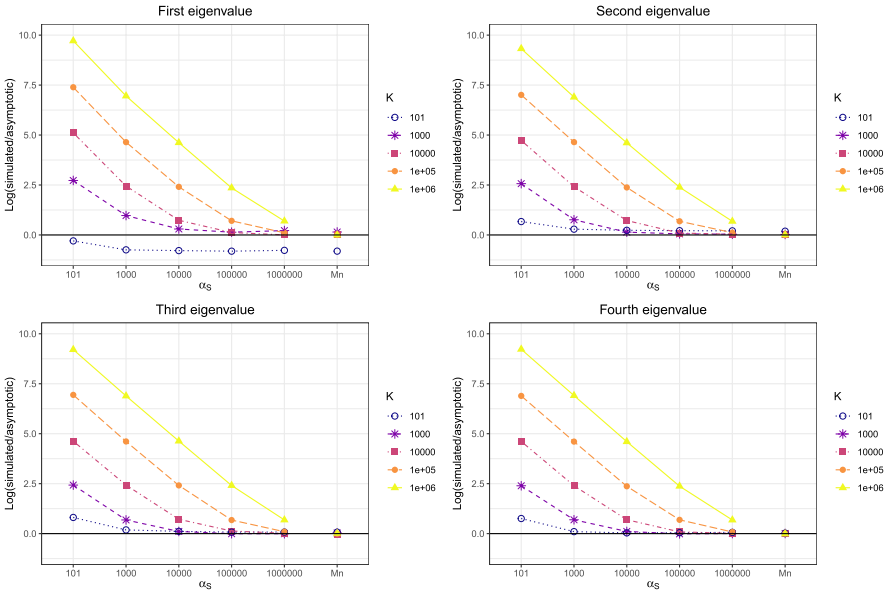

Fig. 7 Log-ratios of the eigenvalues under finite K and α_S, based on Monte Carlo simulation, and eigenvalues of the variance-covariance matrix of the normal approximation based on multinomial counts under multinomial (Mn) and Dirichlet-multinomial (connected dots) counts. Value 0 means means perfect correspondence between the simulated and asymptotic values. Note that the scale differs from that in Fig. 5

nate appears clearly worse under Bayesian replacement compared to imputing zeroes with 0.5, exhibiting heavier tails (Figs. 3 and S3). This is due to the fact that counts imputed by Bayesian replacement tend to be smaller than the fixed pseudo-counts of 0.5, resulting in distinctly lower values of the ilr coordinates for compositions involving zero-count observations. The last ilr coordinate, based on the two most common classes with no zero-count observations, had a similar distribution between the two approaches.

5.4.3 Variable total count

We next summarise the performance of approximation (12) under a lognormally-distributed total count (data generating scenarios (c) and (d)). The variability of the total count (σ^2) is evident in the illustration of the composition of the counts (Fig. S6), but after scaling the counts into proportions, the effect of variation in the total count is barely distinguishable (Fig. S7).

Based on the Q–Q-plots, the distributions of the proportions are quite similar under both fixed and varying total counts (Fig. S8). Importantly, when either K or α_S is small, the distribution of the first proportion is clearly skewed under both fixed and variable total counts (Fig. S8). When the expected value of the total count is small or the proportions do not exhibit extra-multinomial variation, the distribution of the fifth, i.e. the most common proportion tends to have heavier tails under variable total

count, leading to observations with excess deviation from the mean. This tailedness also follows into the distributions of the ilr coordinates (Fig. S9).

The approximation for the expected values (Fig. S10) performs nearly equally well under overdispersion in the total count as it does under fixed total count (cf. Figure 4). However, when μ is small, the approximation is slightly worse for the first two coordinates including also rarer classes. The approximation for the eigenvalues (Fig. S11) of the variance-covariance matrix also works nearly equally well (cf. Figure 5).

5.4.4 Rare counts and the choice of contrast matrix

We investigated how the rarity or commonness of the classes and the choice of the contrast matrix affect the performance of approximation (11). First, we investigated a scenario with a pivotal contrast matrix when the expected class probabilities are (0.50, 0.30, 0.15, 0.04, 0.01). With this choice, the first ilr coordinate contrasts the most abundant class against the other classes and the last ilr coordinate involves the two rarest classes. For each coordinate, the asymptotic expected values now deviate from the simulated values more than with the previous choice of expected class probabilities, (0.01, 0.04, 0.15, 0.30, 0.50), when the total count is small (Fig. S12). This stems from the rare classes now being included in all ilr coordinates and the log-ratio transformation being sensitive to small values. The approximation is poorest for the fourth ilr coordinate (bias up to 40 %).

Second, we investigated a scenario with homogeneous expected class probabilities (0.20, 0.20, 0.20, 0.20, 0.20). In terms of both the expected value and the eigenvalues, the approximation works better than under different class probabilities, even under sparse counts (Fig. S13 and Fig. S14).

5.4.5 Multimodal proportions

Finally, we investigated the performance of approximation (11) when the distribution of proportions and the ensuing counts does not have a mode, i.e., when some or all of the Dirichlet parameters α_j are < 1. We set $\boldsymbol{\alpha} = (0.01, 0.04, 0.15, 0.30, 0.50)$(with $\alpha_S = 1$) and $\boldsymbol{\alpha} = (0.1, 0.4, 1.5, 3, 5)$(with $\alpha_S = 10$). With the first choice, the multimodality of proportions is evident, as there are samples that nearly exclusively consist of observations in just one class. In general, the compositions of counts exhibit a considerable amount of heterogeneity (Fig. S15). Based on Q–Q-plots, it is clear that the distribution of the coordinates clearly deviates from the normal distribution, especially when all components of α are less than 1 (Fig. S16). The excess sparsity affects the performance of the normal approximation, the simulated expected values and eigenvalues of variance-covariance matrices clearly deviating from the asymptotic values (Fig. S17 and Fig. S18).

6 Discussion

An asymptotic normal approximation to the distribution of the ilr coordinates when based on purely multinomial counts with fixed class probabilities, valid with large enough total count, has been derived before (Graffelman et al. 2015). Here, we investigated conditions for the asymptotic normality of ilr coordinates when based on counts with a compound multinomial distribution, i.e., when the multinomial probabilities are not fixed but follow some mixing distribution. We showed that the asymptotic normality of counts and the resulting ilr coordinates holds when either the mixing distribution converges to a constant or when it admits a suitable limiting distribution. In the special case of a unimodal Dirichlet as the mixing distribution, the asymptotic normality holds when the total count approaches infinity and the variability of the class-specific probabilities across the population goes to zero, irrespective of which of the two converges faster.

We derived an explicit expression for the normal approximation of the ilr coordinates under Dirichlet-multinomial counts. Our simulation study confirmed the performance of the approximation with large enough total count K and moderate extra-multinomial variability. Importantly, with large enough K the distribution of the proportions can exhibit more extra-multinomial variation without compromising the satisfactory behaviour of the approximation, while under smaller K the distribution of the proportions needs to be closer to multinomial in order to the approximation to perform well. The variability in the class probabilities (regulated by the sparsity parameter α_K) affects both the expected value and variance-covariance matrices of the ilr coordinates. When comparing our approximation to an approximation relying on the assumption of multinomial counts, we observed superior performance especially with regard to the eigenvalues of the variance-covariance matrix of the coordinates. Of note, adding variability to the total count did not largely affect the performance of the approximation, reflecting the scale invariance of compositional data. Even though in some scenarios it induced heavy tails to the distributions of the coordinates, the performance of the approximation was otherwise good with varying K.

In microbiome data, the read depth determines the total microbial count but is arbitrary and varies across samples. In such situations, the total count is not suitable as the basis for statistical modelling. Instead, the analysis should be based on relative counts whereby compositional data analysis and models built on ilr coordinates are a viable option (Gloor et al. 2017). We chose a general multinomial compound distribution to induce overdispersion to the counts and presented conditions under which the resulting ilr coordinates could be modelled using the normal distribution.

In compositional analysis of microbiome data, the normality of log-ratio transformations is commonly assumed (Sohn and Li 2019; Zhang et al. 2021). However, compositional count observations on the microbiome often exhibit extreme extra-multinomial dispersion (sparsity), i.e. the distribution of counts is not unimodal. Assuming normality of the ilr coordinates may then be ill-justified and, if falsely presumed, may lead to biased or inefficient estimation and false conclusions in mediation analysis as well as other data analytical tasks. In practice, it may be advisable to evaluate the extent of extra-multinomial dispersion in the exploratory phase of the

analysis [(in the context of Dirichlet distribution, see e.g. (Minka 2000)]. If statistical modelling is based on assuming normality, taxa which have a non-unimodal distribution should be discarded. A downside of restricting the analysis to taxa whose distribution shows reasonable homogeneity across samples is that potentially interesting information may be lost if a large number of classes are omitted and the analysis only concerns a subcomposition, i.e., a subset of the original composition. Alternatively, in the context of microbiome data, one may have to investigate only higher taxonomic levels that may be less sparse or to use aggregations, i.e., add up specific parts of a composition. Such aggregations can be based on e.g. taxonomic knowledge or be defined in a data-driven manner with respect to certain response (Gordon-Rodriguez et al. 2021). Furthermore, the taxa can be grouped based on their interrelationships revealed by hierarchical clustering (Boyraz et al. 2022). Even when the data are unimodal, the normality of the ilr coordinates did not hold for all coordinates under small K and small α_S. Assuming normality when modelling ilr coordinates as dependent variables could thus lead to biased or inefficient estimation due to the too heavy tails of the distribution. Furthermore, if the moments of the asymptotic approximation are relied on in the statistical analysis of ilr coordinates with small K and α_S, the regression coefficients may be subject to upward bias while their standard errors could be overestimated (Fig. 5, first eigenvalue, $K = 101$) or underestimated (Fig. 5, first eigenvalue, $\alpha_S = 101$ and $K > 101$), thus leading to too wide or too narrow confidence intervals, respectively, and hence to increased type II and type I errors.

Based on our results, even under unimodal counts it may be meaningful to build the contrasts in such a manner that the rarest classes are not involved in all of the coordinates, given that the empirical research question allows this choice. This advice obviously does not apply if, instead of the ilr coordinates, one uses centered log-ratio transformations or other methods that rely on using the whole composition in each ensuing coordinate. Of note, if very sparse taxa are of specific interest, using logistic regression to analyse the presence/absence of such taxa (i.e., dichotomising the taxa based on the detection limit) could provide meaningful insights (Pelto et al. 2025).

Microbiome data are often very high-dimensional, while we here have investigated the distribution of the ilr coordinates in low dimensions only. Nevertheless, we surmise that the asymptotic normality as shown here holds regardless of the dimension, even when the total count is variable. Of note, our treatment of the normal approximation is valid for general ilr coordinates as we proved their asymptotic normality for an arbitrary contrast matrix. However, the choice of the contrast matrix may affect the finite-sample accuracy of the result. Also the rarity of specific classes and their position in the contrast matrix, i.e., their role in the coordinates, may affect the performance of the approximation.

The use of the Dirichlet distribution to induce overdispersion in multivariate count data has been criticised due to the ensuing spurious correlations that might be unrealistic for real data on microbiome: the correlation between the components is negative, whereas the ratios between different components are statistically independent (Comas-Cufí et al. 2020; Mateu-Figueras et al. 2013). As an alternative, the logratio-multinomial-normal distribution has been proposed in model selection and longitudinal microbiome dynamics (Xia et al. 2013; Comas-Cufí et al. 2020; Silverman et al. 2018). Also this approach is a compound multinomial distribution but the

mixing distribution is obtained as the inverse transformation of the ilr coordinates, which are assumed to follow a normal distribution. Importantly, our aim was not to develop a new statistical model for overdispersed count data under constraints on the total count. Rather, our objective was to address a situation where the total count is not informative and as such is unsuitable for statistical modelling. The motivation to use the Dirichlet-multinomial in the simulation study was due to its transparency and the ease of interpretation under this model.

In addition to sequencing data, such as that on the microbiome, the framework of compositional data analysis is applicable in such research questions where the data arise as non-negative observations that can be scaled by their total (Bacon-Shone 2011). One recent example within the field of health sciences considers partition of daily time use (Pasanen et al. 2022). Another example of measurements that could be considered as compositions is dietary intake of macro-nutrients when characterised as percentage of total energy. In addition, compositional data analysis and our findings may be applicable in geology (e.g. mineral composition in a sample) or social sciences (e.g. election polls). Compared to data on the microbiome, compositions based on these kind of data often exhibit less sparsity and less variability in the total count and the normal approximation for the ilr coordinates may thus perform adequately more often than for the microbiome.

Appendix A Proofs

Notation. Denote the J-part unit simplex by $\mathcal{S}^{J-1}$. Let $\boldsymbol{x}_K \mid \boldsymbol{\pi}_K \sim \text{Multinomial}(K, \boldsymbol{\pi}_K)$, where the random vector $\boldsymbol{\pi}_K$ satisfies $\boldsymbol{\pi}_K \in \mathcal{S}^{J-1}$ and

$$\boldsymbol{\pi}_K = (\pi_{K1}, \ldots, \pi_{KJ})' = \boldsymbol{\alpha} + \boldsymbol{z}_K,$$

where $\boldsymbol{\alpha} \in \mathcal{S}^{J-1}$ is fixed and $\boldsymbol{z}_K = (z_{K1}, \ldots, z_{KJ})'$ is such a random vector in $\mathbf{R}^J$ that the sum $\boldsymbol{\alpha} + \boldsymbol{z}_K$ stays in the simplex $\mathcal{S}^{J-1}$. Let $\text{diag}(\boldsymbol{\pi}_K)$ denote a diagonal matrix with elements of $\boldsymbol{\pi}_K$ on the diagonal. Any $h \in \mathcal{S}^{J-1}$ induces a partitioning of [0, 1] into the intervals $[0, h_1], (h_1, h_1 + h_2], \ldots, (\sum_{\ell=1}^{J-1} h_\ell, 1]$. We denote the jth interval in this partitioning by $\mathcal{P}_j(h)$. Finally, we denote the Lebesgue measure of a subset $\mathcal{A}$ of the real line by $\lambda(\mathcal{A})$.

Proof of Theorem 1 We first observe that $\boldsymbol{x}_K$, conditionally on $\boldsymbol{\pi}_K$, has the same distribution as the vector $\boldsymbol{t}_K = (t_{K1}, \ldots, t_{KJ})'$, where

$$t_{Kj} := \sum_{k=1}^{K} \mathbb{I}\left(u_k \in \mathcal{P}_j(\boldsymbol{\pi}_K)\right)$$

and u_k are i.i.d. $U(0, 1)$-variates independent of $\boldsymbol{\pi}_K$ and $\mathbb{I}$ denotes the indicator function. The indicators above then satisfy,

$$\mathbb{I}(u_k \in \mathcal{P}_j(\boldsymbol{\pi}_K)) - \mathbb{I}(u_k \in \mathcal{P}_j(\boldsymbol{\alpha})) = \quad \mathbb{I}(u_k \in \mathcal{P}_j(\boldsymbol{\pi}_K) \setminus \mathcal{P}_j(\boldsymbol{\alpha})) - \mathbb{I}(u_k \in \mathcal{P}_j(\boldsymbol{\alpha}) \setminus \mathcal{P}_j(\boldsymbol{\pi}_K)).$$

Thus the jth component of $\boldsymbol{t}_K$ satisfies

$$\begin{aligned}\frac{1}{\sqrt{K}}(t_{Kj} - K\alpha_j) =& \frac{1}{\sqrt{K}}\left(\sum_{k=1}^{K}\mathbb{I}(u_k \in \mathcal{P}_j(\boldsymbol{\alpha})) - K\alpha_j\right) \\ &+ \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{I}(u_k \in \mathcal{P}_j(\boldsymbol{\pi}_K) \setminus \mathcal{P}_j(\boldsymbol{\alpha})) \\ &- \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{I}(u_k \in \mathcal{P}_j(\boldsymbol{\alpha}) \setminus \mathcal{P}_j(\boldsymbol{\pi}_K)).\end{aligned} \tag{A1}$$

We next show that the second term on the RHS of (A1) is negligible in probability. Denoting the second term by S_K, we observe that, conditional on $\boldsymbol{z}_K$,

$$\sqrt{K}S_K \mid \boldsymbol{z}_K \sim \mathrm{Bin}(K, \lambda(\mathcal{P}_j(\boldsymbol{\pi}_K) \setminus \mathcal{P}_j(\boldsymbol{\alpha}))).$$

Noting further that

$$\lambda(\mathcal{P}_j(\boldsymbol{\pi}_K) \setminus \mathcal{P}_j(\boldsymbol{\alpha})) \leq \left|\sum_{\ell=0}^{j-1}\pi_{K\ell} - \sum_{\ell=0}^{j-1}\alpha_\ell\right| + \left|\sum_{\ell=0}^{j}\pi_{K\ell} - \sum_{\ell=0}^{j}\alpha_\ell\right| \leq 2\sqrt{J}\|\boldsymbol{z}_K\|,$$

where we define $\pi_{K0} = \alpha_0 := 0$, we get

$$\mathrm{E}(S_K) = \frac{1}{\sqrt{K}}\mathrm{E}\{\mathrm{E}(\sqrt{K}S_K \mid \boldsymbol{z}_K)\}) \leq 2\sqrt{KJ}\mathrm{E}\|\boldsymbol{z}_K\|,$$

which is of the order $o(1)$ by our assumptions since $\{\sqrt{K}\mathrm{E}\|\boldsymbol{z}_K\|\}^2 \leq K\mathrm{E}(\|\boldsymbol{z}_K\|^2)$. Similarly, we get for the variance that

$$\begin{aligned}\mathrm{Var}(S_K) =& K\mathrm{Var}(\lambda_K) + \mathrm{E}\{\lambda_K(1-\lambda_K)\} \\ =& (K-1)\mathrm{E}(\lambda_K^2) - K\{\mathrm{E}(\lambda_K)\}^2 + \mathrm{E}(\lambda_K).\end{aligned}$$

where $\lambda_K := \lambda(\mathcal{P}_j(\boldsymbol{\pi}_K) \setminus \mathcal{P}_j(\boldsymbol{\alpha}))$. Arguing as before, we see that the last two terms above are of the order $o(1)$. For the first term, we have

$$(K-1)\mathrm{E}(\lambda_K^2) \leq 4KJ\mathrm{E}\|\boldsymbol{z}_K\|^2,$$

which is $o(1)$ by our assumptions.

Thus the second term on the RHS of (A1) is negligible in probability. Similarly one can show that also the third term is of the order $o_p(1)$. Thus the joint limiting distribution of the elements of t_K is determined solely by the respective first terms on the RHS of (A1). The claim now follows from the standard central limit theorem. □

Proof of Theorem 2 We first decompose as follows,

$$\frac{\sqrt{\alpha_K}}{K}(\boldsymbol{x}_K - K\boldsymbol{\alpha}) = \frac{\sqrt{\alpha_K}}{K}(\boldsymbol{x}_K - K\boldsymbol{\pi}_K) + \sqrt{\alpha_K}(\boldsymbol{\pi}_K - \boldsymbol{\alpha}),$$

where the second term has the limiting distribution $\mathcal{D}$. Hence, it remains to show that the first term is negligible in probability. We achieve this by establishing that its first two moments vanish when $K \to \infty$.

By the law of total expectation (conditioning on $\boldsymbol{\pi}_K$),

$$\frac{\sqrt{\alpha_K}}{K}\mathrm{E}(\boldsymbol{x}_K - K\boldsymbol{\pi}_K) = \frac{\sqrt{\alpha_K}}{K}\mathrm{E}(K\boldsymbol{\pi}_K - K\boldsymbol{\pi}_K) = 0.$$

Similarly, by the law of total covariance,

$$\begin{aligned}
&\mathrm{Cov}\left\{\frac{\sqrt{\alpha_K}}{K}(\boldsymbol{x}_K - K\boldsymbol{\pi}_K)\right\} \\
=&\frac{\alpha_K}{K^2}\left[\mathrm{Cov}\{\mathrm{E}(\boldsymbol{x}_K - K\boldsymbol{\pi}_K \mid \boldsymbol{\pi}_K)\} + \mathrm{E}\{\mathrm{Cov}(\boldsymbol{x}_K - K\boldsymbol{\pi}_K \mid \boldsymbol{\pi}_K)\}\right] \\
=&\frac{\alpha_K}{K^2}K\mathrm{E}\{\mathrm{diag}(\boldsymbol{\pi}_K) - \boldsymbol{\pi}_K\boldsymbol{\pi}_K'\}.
\end{aligned}$$

The resulting expectation term is bounded as can be seen by writing,

$$\begin{aligned}
\|\mathrm{E}\{\mathrm{diag}(\boldsymbol{\pi}_K) - \boldsymbol{\pi}_K\boldsymbol{\pi}_K'\}\|_1 \leq&\mathrm{E}\|\mathrm{diag}(\boldsymbol{\pi}_K) - \boldsymbol{\pi}_K\boldsymbol{\pi}_K'\|_1 \\
\leq&\mathrm{E}\|\mathrm{diag}(\boldsymbol{\pi}_K)\|_1 + \mathrm{E}\|\boldsymbol{\pi}_K\boldsymbol{\pi}_K'\|_1 \\
\leq&J + J^2,
\end{aligned}$$

where $\|\cdot\|_1$ is the element-wise ℓ_1-norm. The claim now follows. □

Proof of Corollary 1 The expectation of $\boldsymbol{\pi}_K$ is simply $\tilde{\boldsymbol{\alpha}}$ and its covariance matrix is

$$\frac{1}{\alpha_K + 1}\{\mathrm{diag}(\tilde{\boldsymbol{\alpha}}) - \tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\alpha}}'\}.$$

We then have for $\boldsymbol{z}_K = \boldsymbol{\pi}_K - \tilde{\boldsymbol{\alpha}}$ that

$$\mathrm{E}(\|\boldsymbol{z}_K\|^2) = \mathrm{E}\{\boldsymbol{\pi}_K - \mathrm{E}(\boldsymbol{\pi}_K)\}'\{\boldsymbol{\pi}_K - \mathrm{E}(\boldsymbol{\pi}_K)\} = \mathrm{tr}\{\mathrm{Cov}(\boldsymbol{\pi}_K)\} = \frac{1}{\alpha_K + 1}(1 - \|\tilde{\boldsymbol{\alpha}}\|^2),$$

from which the claim now follows by invoking Theorem 1. □

Proof of Corollary 2 By Theorems 4.2 and 4.3 in Geyer and Meeden (2013), the random vector $\boldsymbol{\pi}_K$ satisfies,

$$\sqrt{\alpha_K}\left(\boldsymbol{\pi}_K - \frac{\alpha_K\tilde{\boldsymbol{\alpha}} - \mathbf{1}_J}{\alpha_K - J}\right) \rightsquigarrow \mathcal{N}(\mathbf{0}, \operatorname{diag}(\tilde{\boldsymbol{\alpha}}) - \tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\alpha}}'),$$

where $\mathbf{1}_J$ is a vector of ones. This implies that

$$\begin{aligned}\sqrt{\alpha_K + 1}(\boldsymbol{\pi}_K - \tilde{\boldsymbol{\alpha}}) &= \sqrt{\frac{\alpha_K + 1}{\alpha_K}}\left(\sqrt{\alpha_K}\left(\boldsymbol{\pi}_K - \frac{\alpha_K\tilde{\boldsymbol{\alpha}} - \mathbf{1}_J}{\alpha_K - J}\right) + \sqrt{\alpha_K}\left(\frac{\alpha_K\tilde{\boldsymbol{\alpha}} - \mathbf{1}_J}{\alpha_K - J} - \tilde{\boldsymbol{\alpha}}\right)\right) \\ &= \sqrt{\frac{\alpha_K + 1}{\alpha_K}}\left(\sqrt{\alpha_K}\left(\boldsymbol{\pi}_K - \frac{\alpha_K\tilde{\boldsymbol{\alpha}} - \mathbf{1}_J}{\alpha_K - J}\right) + \sqrt{\alpha_K}\frac{J\tilde{\boldsymbol{\alpha}} - \mathbf{1}_J}{\alpha_K - J}\right) \\ &\rightsquigarrow \mathcal{N}(0, \operatorname{diag}(\tilde{\boldsymbol{\alpha}}) - \tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\alpha}}').\end{aligned}$$

The claim now follows by invoking Theorem 2. □

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00362-025-01732-8.

Acknowledgements The authors are grateful for the two anonymous reviewers whose comments greatly helped improve the quality and presentation of the manuscript.

Author Contributions NK and KA conceived the research idea and wrote the manuscript. NK performed the simulation study. JV developed the theoretical proofs. JN critically reviewed the paper and contributed to the interpretation of the results. KA, JN and OR supervised the manuscript. All authors discussed the research and provided critical feedback that helped shape the manuscript.

Funding Open Access funding provided by University of Turku (including Turku University Central Hospital). Open Access funding provided by University of Turku (including Turku University Central Hospital). NK has been financially supported by Emil Aaltonen Foundation, Alfred Kordelin Foundation, Finnish Cultural Foundation and the MATTI programme in The University of Turku Graduate School (UTUGS). The work of JV was supported by Research Council of Finland, Grants 347501, 353769 and 368494.

Data Availability Not applicable.

Code Availability The code for the simulation study is available from the corresponding author upon request.

Materials Availability Not applicable.

Declarations

Conflict of interest The authors declare no conflict of interest or financial conflict of interest to disclose.

Ethics approval and consent to participate Not applicable.

Consent for publication All authors have seen and approved the final version of the manuscript and consent for its publication.

References

Aitchison J (1982) The statistical analysis of compositional data. J Roy Stat Soc: Ser B (Methodol) 44(2):139–177

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

Aatsinki A-K, Keskitalo A, Laitinen V, Munukka E, Uusitupa H-M, Lahti L, Kortesluoma S, Mustonen P, Rodrigues AJ, Coimbra B, Huovinen P, Karlsson H, Karlsson L (2020) Maternal prenatal psychological distress and hair cortisol levels associate with infant fecal microbiota composition at 2.5 months of age. Psychoneuroendocrinology 119:104754. https://doi.org/10.1016/j.psyneuen.2020.104754

Aatsinki A-K, Lahti L, Uusitupa H-M, Munukka E, Keskitalo A, Nolvi S, O'Mahony S, Pietilä S, Elo LL, Eerola E, Karlsson H, Karlsson L (2019) Gut microbiota composition is associated with temperament traits in infants. Brain Behav Immun 80:849–858. https://doi.org/10.1016/j.bbi.2019.05.035

Aatsinki A-K, Uusitupa H-M, Munukka E, Pesonen H, Rintala A, Pietilä S, Lahti L, Eerola E, Karlsson L, Karlsson H (2018) Gut microbiota composition in mid-pregnancy is associated with gestational weight gain but not prepregnancy body mass index. J Womens Health 27(10):1293–1301. https://doi.org/10.1089/jwh.2017.6488

Boyraz A, Pawlowsky-Glahn V, Egozcue JJ, Acar AC (2022) Principal microbial groups: compositional alternative to phylogenetic grouping of microbiome data. Brief Bioinform 23(5):328. https://doi.org/10.1093/bib/bbac328

Bacon-Shone J (2011) A short history of compositional data analysis. In: Pawlowsky-Glahn V, Buccianti A (eds) Compositional data analysis: theory and applications, 1st edn. Wiley, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Comas-Cufí M, Martín-Fernández JA, Mateu-Figueras G, Palarea-Albaladejo J (2020) Modelling count data using the logratio-normal-multinomial distribution. SORT-Stat Oper Res Trans 44(1):99–126. https://doi.org/10.2436/20.8080.02.96

Egozcue JJ, Graffelman J, Ortego MI, Pawlowsky-Glahn V (2020) Some thoughts on counts in sequencing studies. NAR Genom Bioinform 2(4):094. https://doi.org/10.1093/nargab/lqaa094

Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003) Isometric logratio transformations for compositional data analysis. Math Geol 35:279–300. https://doi.org/10.1023/A:1023818214614

Fry JM, Fry TR, McLaren KR (2000) Compositional data analysis and zeros in micro data. Appl Econ 32(8):953–959. https://doi.org/10.1080/000368400322002

Filzmoser P, Hron K, Templ M (2018) Geometrical properties of compositional data. Applied compositional data analysis. Springer, Cham, pp 35–68

Fernandes AD, Macklaim JM, Linn TG, Reid G, Gloor GB (2013) ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. PLoS ONE. https://doi.org/10.1371/journal.pone.0067019

Geyer C, Meeden G (2013) Asymptotics for constrained dirichlet distributions. Bayesian Anal 8(1):89–110. https://doi.org/10.1214/13-BA804

Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ (2017) Microbiome datasets are compositional: And this is not optional. Front Microbiol 8:1–6. https://doi.org/10.3389/fmicb.2017.02224

Graffelman J, Ortego MI, Egozcue JJ (2015) On the asymptotic distribution of proportions of multinomial count data. Proceedings of the 6th International Workshop on Compositional Data Analysis S

Graffelman J (2011) Statistical inference for Hardy–Weinberg equilibrium using log-ratio coordinates. Proceedings of CoDaWork'11: 4th international workshop on Compositional Data Analysis, Egozcue JJ, Tolosana-Delgado R, Ortego MI (eds) 2011

Gordon-Rodriguez E, Quinn TP, Cunningham JP (2021) Learning sparse log-ratios for high-throughput sequencing data. bioRxiv. https://doi.org/10.1101/2021.02.11.430695

Keskitalo A, Aatsinki A-K, Kortesluoma S, Pelto J, Korhonen L, Lahti L, Lukkarinen M, Munukka E, Karlsson H, Karlsson L (2021) Gut microbiota diversity but not composition is related to saliva cortisol stress response at the age of 2.5 months. Stress 24(5):551–560. https://doi.org/10.1080/10253890.2021.1895110

Martín-Fernández JA, Barceló-Vidal C, Pawlowsky-Glahn V (2003) Dealing with zeros and missing values in compositional data sets using nonparametric imputation. Math Geol 35:253–278. https://doi.org/10.1023/A:1023866030544

Martín-Fernández J-A, Hron K, Templ M, Filzmoser P, Palarea-Albaladejo J (2014) Bayesian-multiplicative treatment of count zeros in compositional data sets. Stat Model 15(2):134–158. https://doi.org/10.1177/1471082X14535524

Mateu-Figueras G, Pawlowsky-Glahn V, Egozcue JJ (2011) The principle of working on coordinates. In: Pawlowsky-Glahn V, Buccianti A (eds) Compositional data analysis: theory and applications, 1st edn. Wiley, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Mateu-Figueras G, Pawlowsky-Glahn V, Egozcue J-J (2013) The normal distribution in some constrained sample spaces. SORT-Stat Oper Res Trans 37(1):29–56

Minka TP (2000) Estimating a Dirichlet distribution. https://tminka.gitmischub.io/papers/dirichlet/minka-dirichlet.pdf

Pelto J, Auranen K, Kujala JV, Lahti L (2025) Elementary methods provide more replicable results in microbial differential abundance analysis. Brief Bioinform 26(2):130. https://doi.org/10.1093/bib/bbaf130

Palarea-Albaladejo J, Martín-Fernández JA (2015) zcompositions - R package for multivariate imputation of left-censored data under a compositional approach. Chemom Intell Lab Syst 143:85–96. https://doi.org/10.1016/j.chemolab.2015.02.019

Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015) Coordinate representation. In: Modelling and Analysis of Compositional Data, pp. 32–64. Wiley, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Pasanen J, Leskinen T, Suorsa K, Pulakka A, Virta J, Auranen K, Stenholm S (2022) Effects of physical activity intervention on 24-h movement behaviors: a compositional data analysis. Sci Rep 12:8712. https://doi.org/10.1038/s41598-022-12715-2

Silverman JD, Durand HK, Bloom RJ, Mukherjee S, David LA (2018) Dynamic linear models guide design and analysis of microbiota studies within artificial human guts. Microbiome 6:202. https://doi.org/10.1186/s40168-018-0584-3

Sohn MB, Li H (2019) Compositional mediation analysis for microbiome studies. Annals Appl Stat 13(1):661–681. https://doi.org/10.1214/18-AOAS1210

Wang C, Hu J, Blaser MJ, Li H (2020) Estimating and testing the microbial causal mediation effect with high-dimensional and compositional microbiome data. Bioinformatics 36(2):347–355. https://doi.org/10.1093/bioinformatics/btz565

Xia F, Chen J, Fung WK, Li H (2013) A Logistic Normal Multinomial Regression Model for Microbiome Compositional Data Analysis. Biometrics 69(4):1053–1063. https://doi.org/10.1111/biom.12079

Zhang H, Chen J, Li Z, Liu L (2021) Testing for mediation effect with application to human microbiome Data. Stat Biosci 13:313–328. https://doi.org/10.1007/s12561-019-09253-3

Authors and Affiliations

Noora Kartiosuo[1,2,3,4] · Joni Virta[1] · Jaakko Nevalainen[5] · Olli Raitakari[2,3,6] · Kari Auranen[1,7]

✉ Noora Kartiosuo
noora.kartiosuo@utu.fi

Joni Virta
joni.virta@utu.fi

Jaakko Nevalainen
jaakko.nevalainen@tuni.fi

Olli Raitakari
olli.raitakari@utu.fi

Kari Auranen
kari.auranen@utu.fi

[1] Department of Mathematics and Statistics, University of Turku, 20014 Turku, Finland

[2] Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, 20014 Turku, Finland

[3] Centre for Population Health Research, University of Turku, 20014 Turku, Finland

[4] Murdoch Children's Research Institute, 3052 Parkville, Victoria, Australia

[5] Health Sciences Unit, Faculty of Social Sciences, Tampere University, 33014 Tampere, Finland

[6] Department of Clinical Physiology and Nuclear Medicine, University of Turku and Turku University Hospital, 20014 Turku, Finland

[7] Department of Clinical Medicine, University of Turku, 20014 Turku, Finland