

PAPER

# Proportionality-based association metrics in count compositional data

Kevin McGregor,<sup>1,\*</sup> Nneka Okaeme,<sup>1,2</sup> Reihane Khorasaniha,<sup>3</sup> Simona Veniamin,<sup>4</sup> Juan Jovel,<sup>5</sup> Richard Miller,<sup>3</sup> Ramsha Mahmood,<sup>3</sup> Morag Graham,<sup>3,6</sup> Christine Bonner,<sup>6</sup> Charles N. Bernstein,<sup>3</sup> Douglas L. Arnold,<sup>7</sup> Amit Bar-Or,<sup>8</sup> Janace Hart,<sup>9</sup> Ruth Ann Marrie,<sup>3</sup> Julia O'Mahony,<sup>10</sup> E. Ann Yeh<sup>11</sup>, Yinshan Zhao,<sup>12</sup> Brenda Banwell,<sup>8</sup> Emmanuelle Waubant,<sup>9</sup> Natalie Knox,<sup>3,6</sup> Gary Van Domselaar,<sup>3,6</sup> Feng Zhu,<sup>12</sup> Ali I. Mirza,<sup>12</sup> Helen Tremlett<sup>12</sup> and Heather Armstrong<sup>3,4</sup>

<sup>1</sup>Department of Mathematics and Statistics, York University, Toronto, Ontario, M3J 1P3, <sup>2</sup>Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, N2L 3G1, <sup>3</sup>University of Manitoba, Winnipeg, Manitoba, R3T 2N2, <sup>4</sup>University of Alberta, Edmonton, Alberta, T6G 2R3, <sup>5</sup>University of Calgary, Calgary, Alberta, T2N 1N4, <sup>6</sup>Public Health Agency of Canada, Winnipeg, Manitoba, R3E 3R2, <sup>7</sup>McGill University, Montréal, Québec, H3A 0G4, <sup>8</sup>Children's Hospital of Philadelphia, University of Pennsylvania, Perelman School of Medicine, Philadelphia, Pennsylvania, 19104, <sup>9</sup>Weill Institute for Neurosciences, University of California San Francisco, San Francisco, California, 94158, <sup>10</sup>Mellen Center for Multiple Sclerosis, Cleveland Clinic, Cleveland, Ohio 44195, <sup>11</sup>Division of Neurology, Hospital for Sick Children, Department of Pediatrics, University of Toronto, Toronto, Ontario, M5R 0A3 and <sup>12</sup>University of British Columbia, Vancouver, British Columbia, V6T 1Z4

\*Corresponding author. kevinmccg@yorku.ca

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

## Abstract

**Motivation:** Compositional data comprise vectors that describe the constituent parts of a whole. Data arising from various -omics platforms such as 16S and RNA-sequencing are compositional in nature. However, correlations between features on raw counts have no meaningful interpretation. Metrics of proportionality were formulated to address this problem. However, there is an inherent bias that arises when calculating these metrics empirically on count-based measures due to variability in read depths.

**Results:** We quantify the bias introduced by empirically calculating proportionality-based association metrics in count data. Additionally, we propose a means of estimating these metrics within a logit-normal multinomial model in pursuit of more accurate estimates. The model-based estimates are shown to outperform empirical estimates in simulated data, and are additionally applied to a mouse embryonic stem-cell single-cell sequencing dataset as well as a pediatric-onset multiple sclerosis metagenomic dataset.

**Availability and Implementation:** An R package is available at <https://CRAN.R-project.org/package=countprop>.

**Supplementary information:** Supplementary data are available at Bioinformatics online.

**Key words:** proportionality metrics, compositional data, count data, single-cell sequencing, microbiome

## Introduction

Compositional data comprise vectors of quantitative measures that describe the constituent parts of a whole. The only information in compositional data is found in the *ratios* between the different parts of the vector; the individual measures are inherently uninterpretable [1]. Data arising from various omics platforms are compositional in nature—for instance, RNA-sequencing (RNA-Seq), 16S amplicon-sequencing, and single-cell RNA-sequencing are all examples of platforms that yield compositional data. The primary reason for the compositional classification of these platforms is that the number of sequences varies across samples due to technical

artifacts in the sequencing process, meaning that within-sample counts must either be normalized to a unit sum, or log-ratio transformed in order to effect any kind of meaningful interpretation [9]. Some microbiome studies have relied on *rarefying* such data, which involves resampling the counts to a constant read depth across all samples. However, it has been recognized that such a procedure essentially amounts to discarding valid data and should therefore be discouraged [18].

This interpretability problem is particularly worrisome when considering associations between features—the prime example being Pearson's correlation. Due to varying read depths, correlations on raw counts have no meaningful interpretation.

Similarly, when normalizing the counts to proportions, a negative bias is introduced in the correlations as a result of the unit-sum constraint.

An appropriate class of measures based on *proportionality* has been established to address the issue of quantifying association in compositional data. The most basic such measure is the variation matrix  $\mathbf{V}$  whose elements are defined as follows: suppose  $\mathbf{p} = (p_1, \dots, p_D)$  is a random vector containing proportions such that  $\sum_{j=1}^D p_j = 1$ . The element in row  $j$  and column  $k$  of  $\mathbf{V}$ , which describes the association between features  $j$  and  $k$  is given by:

$$v(p_j, p_k) = \text{var} \left[ \log \left( \frac{p_j}{p_k} \right) \right].$$

When  $p_j$  is exactly proportional to  $p_k$ , we have that  $v(p_j, p_k) = 0$ . Hence, the smaller the value of  $v(p_j, p_k)$  the stronger the association between the abundances of features  $j$  and  $k$ . Note that the elements of  $\mathbf{V}$  are non-negative, and thus the variation matrix does not give information about the direction of the associations of the pairs of features included in the matrix.

One important issue with the variation matrix is that its elements are variances of different pairs of features, meaning that they do not share a common scale. To address this issue, [7] developed alternative measures of proportionality.

The first is called  $\phi$ , and is simply  $v(p_j, p_k)$  scaled by the variance of the logarithm of the first argument. It is defined as:

$$\phi(p_j, p_k) = \frac{v(p_j, p_k)}{\text{var}[\log(p_j)]}, \quad (1)$$

with  $\phi(p_j, p_k) \geq 0$ . The interpretation is the same as for  $v$ , only now it is scaled based on its first argument. The main downside of this metric is that it is not symmetric with regards to its arguments. A final proportionality metric is defined as:

$$\rho(p_j, p_k) = \frac{2 \text{cov}[\log(p_j), \log(p_k)]}{\text{var}[\log(p_j)] + \text{var}[\log(p_k)]}, \quad (2)$$

with  $-1 \leq \rho(p_j, p_k) \leq 1$ . If  $\rho \approx 1$ , then  $p_j$  and  $p_k$  are strongly proportional, and if  $\rho \approx -1$ , then  $p_j$  and  $p_k$  are strongly inversely proportional.

Metrics of proportionality recently gained more recognition, when [25] suggested their use to detect gene-gene associations in single-cell sequencing experiments. This publication compared numerous measures of association across multiple datasets by evaluating their ability to create functionally coherent single-cell gene co-expression networks. There was a clear advantage for the proportionality metrics across the datasets.

The one common property of these kinds of metrics is that they are *scale invariant*, which is the main property allowing their use in compositional data. However, classical techniques for the analysis of compositional data were developed with continuous measures in mind; for example, the constituent parts of a soil sample. Conversely, sequencing platforms provide data in the form of *counts*. A number of recent studies have shed light on problems that arise when applying traditional methods for compositional data analysis on count-based compositional data. [14] showed that proportionality cannot be exactly represented in count-based compositional data, which they referred to as *lattice* compositional data. [15] showed a similar phenomenon in logarithmic transformations in single-cell sequencing; varying read depths among cells led to systematic errors and spurious differences in expression. Relatedly, [6] presented novel modelling techniques in light of these issues in count data.

In addition to the representation problems outlined in [14], we further claim that there is in fact a *bias* that arises when applying proportionality metrics on count-based data, which is a result of added variability from the randomness of the sequencing process. In this paper, we quantify this bias and provide an alternative model-based means of estimating proportionality metrics in count-based compositional data.

## Methods

### Bias

In this section we outline the inherent bias in estimating metrics of proportionality using observed counts. Suppose that we have a matrix  $\mathbf{Y}_{n \times (J+1)}$  containing counts of  $J+1$  features (e.g. genes, species, etc.) from  $n$  samples. Let the entries of  $\mathbf{Y}$  be  $y_{ij}$ , representing the observed count of feature  $j$  in sample  $i$ . We seek to compare the result of plugging counts  $y_{ij}$  (or equivalently observed proportions  $\hat{p}_{ij} = y_{ij}/n_i$ ) into  $v$ ,  $\phi$ , and  $\rho$  against the true (unknown) values which instead take *unobserved* proportions  $p_{ij}$  as arguments to these metrics. In each case, there is no exact closed form for the bias, and we instead use an approximation based on a second-order Taylor series expansion. We summarize the results here, but full derivations can be found in Section S2. In each case, it is clear that the bias is more pronounced when the read depths are small, on average. However, it is also evident that the amount of variation in the read depths itself plays into the bias, with more variability leading to a larger discrepancy between the empirical and true values of the metrics. The underlying distribution of the unobserved proportion vector  $\mathbf{p}_i = (p_{i1}, \dots, p_{i(J+1)})$  is also relevant.

First, we consider  $v(y_{ij}, y_{ik})$ , which can be approximated as:

$$v(y_{ij}, y_{ik}) \approx v(p_{ij}, p_{ik}) + b^*, \quad (3)$$

where,

$$\begin{aligned} b^* &= \frac{1}{4} \text{var} \left[ n_i^{-1} (p_{ij}^{-1} - p_{ik}^{-1}) \right] \\ &\quad - \text{cov} \left[ \log \left( \frac{p_{ij}}{p_{ik}} \right), n_i^{-1} (p_{ij}^{-1} - p_{ik}^{-1}) \right] \\ &\quad + \mathbb{E} \left[ n_i^{-1} \right] \left( \mathbb{E} [p_{ij}^{-1}] + \mathbb{E} [p_{ik}^{-1}] \right). \end{aligned} \quad (4)$$

As expected, the expectation of  $1/n_i$  is an important factor, however, the expectations of  $1/p_{ij}$  and  $1/p_{ik}$  are also present, suggesting that the discrepancy will be more pronounced in less abundant features. The bias for  $\phi$  can be expressed as:

$$\phi(y_{ij}, y_{ik}) \approx \frac{\phi(p_{ij}, p_{ik}) + \frac{b^*}{\text{var}[\log(p_j)]}}{1 + \frac{r_{ij} - \mathbb{E}[n_i^{-1}] (\mathbb{E}[p_{ij}^{-1}] - c_{ijj} - 1)}{\text{var}[\log(p_j)]}}, \quad (5)$$

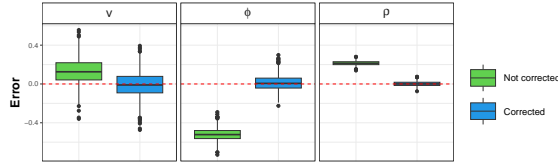
where

$$c_{ijj} = \text{cov} \left[ \log(p_{ij}), p_{ij}^{-1} \right], \quad (6)$$

$$r_{ij} = \text{var} \left[ \log(n_i) + (2n_i)^{-1} \left( 1 - p_{ik}^{-1} \right) \right] \quad (7)$$

Finally, we consider the analogous result for  $\rho$ :

$$\rho(y_{ij}, y_{ik}) \approx \frac{\rho(p_{ij}, p_{ik}) + \frac{b_\rho^*}{\text{var}[\log(p_j)] + \text{var}[\log(p_k)]}}{1 + \frac{r_{ij} + r_{ik} - \mathbb{E}[n_i^{-1}] (2 + c_{ijj} + c_{ikk} - \mathbb{E}[p_{ij}^{-1}] - \mathbb{E}[p_{ik}^{-1}])}{\text{var}[\log(p_j)] + \text{var}[\log(p_k)]}}, \quad (8)$$



**Fig. 1.** Estimation error for a single entry of  $v$ ,  $\phi$  and  $\rho$  calculated using observed counts in 2000 simulation replications. The green is not corrected for bias, the blue is corrected for bias using Equations 3, 5, and 8.

where,

$$b_{\rho}^* = 2 \text{var}[\log(n_i)] - \left(2 - \mathbb{E}[p_{ij}^{-1}] - \mathbb{E}[p_{ik}^{-1}]\right) \text{cov}[\log(n_i), n_i^{-1}] + \frac{1}{2} q_{ijk} - \mathbb{E}[n_i^{-1}] (c_{ijk} + c_{ikj} - 2)$$

and,

$$c_{ijk} = \text{cov}[\log(p_{ij}), p_{ik}^{-1}], \\ q_{ijk} = \text{cov}\left[n_i^{-1} (1 - p_{ij}^{-1}), n_i^{-1} (1 - p_{ik}^{-1})\right]$$

We can once again see that the bias is less pronounced when the mean of the read depths  $n_i$  is large. However, the variation of  $n_i$  as well as distribution of the proportions  $\mathbf{p}_i$  are also important. This suggests that there could still be bias present for estimates for certain feature pairs even when the read depths are large.

To numerically investigate these biases, we run a small simulation from a multinomial logit-normal distribution (see the next section for the definition of this distribution). We simulate  $J + 1 = 7$  features on  $n = 2000$  samples, with 1000 simulation replications. In each replication, the estimated proportionality metrics on the raw counts  $v(y_{ij}, y_{ik})$ ,  $\phi(y_{ij}, y_{ik})$ , and  $\rho(y_{ij}, y_{ik})$  are calculated (after imputing zeros using Bayesian-multiplicative replacement [22], if necessary) and compared to their respective true values  $v(p_{ij}, p_{ik})$ ,  $\phi(p_{ij}, p_{ik})$ , and  $\rho(p_{ij}, p_{ik})$ . Additionally, a bias-corrected version of the estimator is calculated by adjusting the empirical estimates using Equations 3, 5, and 8.

Figure 1 shows the difference between the estimated values and the true values for a single element of the matrix. It is evident that all of  $v$ ,  $\phi$ , and  $\rho$  exhibit strong biases when calculated on the raw counts. In all cases, the sampling distribution of the corrected estimator is approximately centered around zero. Results for the full matrix can be seen in Figures S1-S3.

It should be noted that this bias-correction approach is not feasible in practice, as it would require knowledge of the true population parameters, which happen to be known in this simulation. As an alternative, we propose directly modelling the distribution of the proportions  $\mathbf{p}_i$ . The parameters of this distribution then define the values of  $v$ ,  $\phi$ , and  $\rho$ , allowing the proportionality metrics to be estimated free from the variation induced by the sequencing process itself.

## Model

We proceed in the context of the multinomial logit-normal model, posited by [30]. This model has been extremely popular in statistical modelling of the microbiome [29, 24, 17]. In this formulation, we assume that the count vector for sample  $i$ ,

denoted by  $\mathbf{Y}_i = (y_{i1}, \dots, y_{i(J+1)})$ , is distributed as:

$$\mathbf{Y}_i \sim \text{Multinomial}(n_i, \mathbf{p}_i)$$

where  $J + 1$  is the total number of features observed among all samples and  $n_i = \sum_{j=1}^{J+1} y_{ij}$  is the read depth. The proportion vector for individual  $i$  is assumed to follow the logit-normal distribution, which is characterized by the inverse additive log-ratio (ALR) transformation:

$$\mathbf{p}_i = \text{alr}^{-1}(\mathbf{w}_i) \\ = \mathcal{C}(\exp(w_{i1}), \dots, \exp(w_{iJ}), 1)$$

where  $\mathcal{C}(x_1, \dots, x_k) = (x_1, \dots, x_k) / (\sum_{j=1}^k x_j)$  is the closure operation. Each  $\mathbf{w}_i = (w_{i1}, \dots, w_{iJ})$  is an unobserved, latent vector assumed to be multivariate-normal:

$$\mathbf{w}_i \sim \text{MV-Normal}_J(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \quad (9)$$

We set the  $(J + 1)^{\text{th}}$  feature as the reference feature in the ALR transformation (without loss of generality; any feature could be moved to that column). The read depths themselves are random variables on which we assume a log-normal distribution:

$$n_i \sim \text{Log-Normal}(\mu_{\ell}, \sigma_{\ell}^2)$$

The log-normal distribution is appropriate given that the read depth distribution in a dataset can span multiple orders of magnitude. This assumption has been used in previous literature [16].

To handle the possibility of a large number of features  $J + 1$  relative to the sample size  $n$ , we apply the Graphical Lasso (GLasso) penalty [8] to the multivariate normal log-likelihood for the  $\mathbf{w}_i$  vectors:

$$\ell(\boldsymbol{\Sigma}^{-1} | \mathbf{W}) = \frac{n}{2} \log \det \boldsymbol{\Sigma}^{-1} \\ - \frac{1}{2} \sum_{i=1}^n (\mathbf{w}_i - \boldsymbol{\mu})^{\top} \boldsymbol{\Sigma}^{-1} (\mathbf{w}_i - \boldsymbol{\mu}) - \lambda \|\boldsymbol{\Sigma}^{-1}\|_1$$

where  $\|\cdot\|_1$  denotes the  $L_1$ -norm, and  $\lambda \geq 0$  is the penalty parameter. This penalization is very important in guarding against spurious associations when the number of features is large.

## Estimating proportionality metrics

In the logit-normal framework, the variation matrix elements have a convenient form, namely:

$$v(p_{ij}, p_{ik}) = \begin{cases} \sigma_j^2 + \sigma_k^2 - 2\sigma_{jk} & \text{if } j \neq J+1 \text{ and } k \neq J+1 \\ \sigma_j^2 & \text{if } j \neq J+1 \text{ and } k = J+1 \\ \sigma_k^2 & \text{if } j = J+1 \text{ and } k \neq J+1. \end{cases} \quad (10)$$

For  $\phi$  and  $\rho$  we also need to estimate the covariance matrix of the log-proportions, which we denote by  $\boldsymbol{\Omega} = \text{var}[\log(\mathbf{p})]$ . The relationship between the covariance of the log-proportions is given by  $\boldsymbol{\Sigma} = \mathbf{F}\boldsymbol{\Omega}\mathbf{F}^{\top}$ , where  $\mathbf{F} = [\mathbf{I}_J, -\mathbf{1}_J]$ , with  $\mathbf{I}_J$  representing the  $J \times J$  identity matrix, and  $\mathbf{1}_J$  representing a  $J$ -dimensional vector of ones. In the logit-normal distribution there is no closed-form expression for the matrix  $\boldsymbol{\Omega}$ . The additive log-ratio variance  $\boldsymbol{\Sigma}$  does not itself admit a unique  $\boldsymbol{\Omega}$  [1]; the mean of the distribution is also required to determine  $\boldsymbol{\Omega}$ .

Consequently, to estimate the log-proportion variance  $\Omega$  from  $\mu$  and  $\Sigma$ , we use a second-order Taylor approximation [10] based on the transformation from the  $w_{ij}$  scale to the  $\log(p_{ij})$  scale. We can approximate the covariance of the log-proportion vector as:

$$\Omega \approx \mathbf{Q}_\mu \Sigma \mathbf{Q}_\mu^\top + \frac{1}{2} \text{tr} \left( \Sigma \mathbf{H}_\mu \Sigma \mathbf{H}_\mu \right), \quad (11)$$

where,

$$\mathbf{Q}_\mu = \begin{pmatrix} \mathbf{I}_J \\ 0 \end{pmatrix} - \mathbf{1}_{J+1} [\text{alr}^{-1}(\mu)]^\top$$

$$\mathbf{H}_\mu = \text{alr}^{-1}(\mu) [\text{alr}^{-1}(\mu)]^\top - \text{diag} [\text{alr}^{-1}(\mu)].$$

Estimates of  $\phi$ , and  $\rho$  can then be obtained by plugging the approximations from Equations 10 and 11 into their respective definitions.

### Fitting the model

To fit the logit-normal multinomial model with graphical lasso penalty applied to the ALR-scale precision matrix, we perform maximum likelihood estimation through an implementation of the EM-algorithm posited in [11, 27]. Model selection (through the choice of tuning parameter  $\lambda$ ) is performed using the Extended Bayesian Information Criterion (EBIC) [4]. The steps of the algorithm can be found in Section S4.

### Simulation study

We performed an extensive simulation study to compare the performance of model-based vs. empirical estimates. We focus on  $\rho$  in the simulations since it is the most interpretable metric, and we can fix it to zero for certain feature pairs allowing precision-recall analysis (whereas the null for  $v$  and  $\phi$  is infinity). We also include an additional estimator, which uses empirical estimates of  $\mu$  and  $\Sigma$ , based on simply taking the additive log-ratio of the observed counts and calculating the sample means and variances of the features. These parameter estimates are then inserted into Equation 11. We will refer to this latter estimator as the “plugin” estimator, and does not require running maximum likelihood estimation.

Our simulation parameters are based on an initial multinomial logit-normal model fitted to single-cell RNA-sequencing data from a mouse embryonic stem-cell dataset from Buettner et al. [2]. The model parameters  $\mu$ ,  $\Omega$ ,  $\Sigma$ ,  $\mu_\ell$ , and  $\sigma_\ell$ , are estimated from the data and are then used as the ground truth in the simulations. See Section S5 for more details about the simulation procedures.

Several scenarios are considered. The first scenario uses all the parameter values that estimated from the Buettner dataset. This scenario allows us to compare the estimated values to the true values used in the simulation using the root mean squared error (RMSE).

The second scenario modifies  $\Omega$  to be a sparse matrix, which allows some of the true values of  $\rho(p_{ij}, p_{ik})$  to be zero. This lets us ascertain the ability of the model-based and empirical estimates to find the non-zero elements of  $\rho(p_{ij}, p_{ik})$ . Since neither the model-based nor empirical estimators will set values exactly to zero, a range of thresholds is applied so that estimates whose  $\rho$  value is greater (in absolute value) than the threshold are considered to be estimated as non-zero. This allows us to calculate the area under the precision-recall curve to compare the model-based and empirical estimates.

The third scenario considers model misspecification. The motivation is to simulate values of the  $\mathbf{w}_i$  vectors from a distribution that is non-Gaussian to see how estimates of  $\rho$  are affected by a misspecified model. To do this, we instead simulate  $\mathbf{w}_i$  from a multivariate non-central  $t$ -distribution with mean  $\mu$ , covariance  $\Sigma$ , and degrees of freedom parameter equal to 2.1. This allows the distribution of the  $\mathbf{w}_i$  terms to have heavier tails than the normal distribution as well as skewness. The rest of the simulation procedure is the same as in the second scenario. Precision-recall curves are used to compare performance.

### Data application

To demonstrate the applicability of the proportionality metric estimation technique introduced in this paper, we apply the methods to two datasets—one is a single-cell RNA-sequencing dataset, and the other is a metagenomic dataset.

The first dataset is a mouse embryonic stem-cell (mESC) single-cell RNA-sequencing dataset from [2]. Flow cytometry was used to sort by cell-type, with cells sorted into G1, S, and G2M cell cycle stages. The original dataset contained 96 samples and 38,390 genes. The goal of our analysis is to focus on genes that were previously shown to be differentially expressed among cell-types. In this analysis, we consider 570 genes associated with cell-cycle based on GO annotations; this list was provided in the original publication. We took a further subset of genes by filtering out genes having greater than 20% dropouts in at least one of the cell-cycle stages. There were 303 genes remaining after filtering. This analysis allows investigation of whether gene-gene associations differ between cell-cycle stages in a subset of genes that were previously known to be related to cell-cycle.

The second dataset comprises shotgun metagenomic sequencing information derived from stool samples procured from participants with pediatric-onset (symptom onset < 18 years of age) multiple sclerosis (MS) and unaffected controls [20, 19]. The goal of the study was to determine how suppression of dietary fibre fermentation can induce inflammation in MS. Data were collected through the Canadian Pediatric Demyelinating Disease Network; all participants were under 22 years of age at the time of stool sample procurement. There were 17 MS participants (14 female) and 20 unaffected controls (16 female) who provided a stool sample. For our analyses, we proceed at the genus level; initially there were 617 genera available in the dataset. We filtered out genera with greater than 10% zeros, leaving 296 genera. Due to the very small sample size of this dataset, we took a further subset of the 100 most abundant remaining genera. This facilitated more stable estimates of the model parameters in the small sample size case. The analysis focuses on which genus pairs have differing associations between MS participants and unaffected controls.

## Results

### Simulation results

Results from the first simulation can be seen in Section S6. The goal of this initial simulation is to investigate the estimation accuracy of  $\rho$  and the elements of  $\Omega$ . In Figure S4, the RMSE is shown for empirical and model-based estimation of  $\rho$ . It is clear that there is a substantial improvement in model-based estimates in most cases, especially when the number of features  $J$  is large, which is an important guard against finding spurious correlations. We show similar results for off-diagonal elements



**Fig. 2.** Area under the precision-recall curve of  $\rho$  estimation methods in simulation data. Error bars represent 25th and 75th percentiles over 50 simulation replications. Simulations vary over sparsity level (e.g. proportion of true  $\rho$  values equal to 0) and number of features  $J$ .

of  $\Omega$  in Figure S5. The RMSE patterns for  $\Omega$  are congruent with those of  $\rho$ .

Next we show the results from the simulation scenario considering sparsity among the  $\rho$  values. Results can be seen in Figure 2, where we compare the distributions of the AUPRC for the different estimates. In all scenarios, the model-based estimates outperform the empirical estimates. The difference in performance is especially pronounced when the number of features  $J$  is large. Intriguingly, even the plugin estimates of  $\rho$  outperform the empirical estimates. This is also an attractive option, as it does not require running maximum likelihood estimation.

Finally, we present the results from repeating the AUPRC simulation, but under severe model misspecification (non-central  $t$ -distribution in place of Gaussian). These results can be seen in Figure S6. Though the differences between the AUPRC are more subdued in this case, there is still a clear advantage for the model-based estimator over the empirical estimator. There is, again, a slight advantage for the plugin estimator over the empirical estimator. This is an encouraging result, as it demonstrates the utility of our estimator for  $\rho$  even if the form of the assumed model is not correct.

All of the simulations show strong evidence of the superiority of the model-based estimates over the empirical estimates for  $\rho$  elements. Thus, the model-based estimates can be used to guard against spurious correlations in when considering the proportionality metric  $\rho$ .

## Data application results

In this section, we outline the results from the two data analyses described in the Methods section. First we discuss the results from the murine single-cell RNA-seq dataset. All gene descriptions were obtained from GeneCards [26].

We consider the estimated  $\rho$  values for each of the pairs of the 303 genes retained after the aforementioned filtering steps. Estimates were obtained using both the empirical estimator and the model-based estimator. We compare the top 100 gene-gene associations (in absolute value) detected using the empirical and model-based estimators. In the G1 phase there were 23 gene pairs in common, in G2M phase there were 24 gene pairs in common, and in S phase there was only 1 gene pair in common (Figure S7). Relatedly, the empirical and model-based estimates are compared in Figure S8. This shows the attenuation towards the null for the model-based estimates, which is an important guard against spurious correlations.

These comparisons between the empirical and model-based estimates show how different the results can be using the empirical estimator for  $\rho$  and highlights the importance of correcting the inherent biases in proportionality metrics applied to count data.

The top 100 gene pairs within each cell stage are shown in Tables S4-6. Additionally, Table 1 summarizes the top differences (in absolute value) in  $\rho$  values between phases G1 and G2M phases; more extensive tables showing the top 100 gene pairs for each of the cell phase comparisons are shown in Tables S1-S3.

In the G1 vs. G2M comparison, the largest  $\rho$  difference is seen between genes STAG3 and CINP, with  $\rho_{G1} - \rho_{G2M} = -0.5005$ . STAG3 encodes a protein involved in the regulation of the cohesion of sister chromatids during cell division; CINP is part of the DNA replication complex and binds chromatin in the G1 phase.

In the G1 vs. S comparison (Table S2), the four largest differences all involve the ARHGEF2 gene, which is implicated in Rho-GTPase activation. The genes its  $\rho$  values differ with the most between G1 and S are ATM, STAG2, CHEK1, and ANAPC4, all of whose expression is much more positively correlated with ARHGEF2 in S phase compared to G1 phase.

In the G2M vs. S comparison (Table S3), ARHGEF2 is again involved in several of the top  $\rho$  differences; namely, with ATM, CHEK1, STAG2, and ANAPC4, again with much stronger positive correlation in S phase compared to G1 phase. Another notable difference is for ANAPC4 and PPM1G, which are slightly negatively correlated in G2M phase and positively correlated in S phase. ANAPC4 is involved in promotion of the metaphase-anaphase transition, and PPM1G is related to negative regulation of cell stress response.

Gene 1	Gene 2	$\rho_{G1}$	$\rho_{G2M}$	$\rho_{G1} - \rho_{G2M}$
FANCI	CACUL1	0.0394	-0.2453	0.2847
CLASP1	SEP-09	0.1426	-0.139	0.2816
TFDP1	CDC16	0.2141	-0.0585	0.2726
TET2	EVI5	0.2575	-0.0135	0.271
ANAPC4	GNAI3	0.242	-0.0258	0.2678
ENSA	MELK	0.2088	-0.0562	0.2651
HAUS3	CDC16	0.0278	-0.2355	0.2633
FANCD2	CCNA2	0.1562	-0.106	0.2622
ENSA	KLHL13	0.0052	0.2715	-0.2663
GNAI2	CCNA2	0.0037	0.274	-0.2703
NCAPG2	TXNIP	-0.1767	0.0974	-0.2741
B230120H23RIK	CCNE1	-0.1921	0.0856	-0.2777
STAG3	ANAPC4	-0.1492	0.1296	-0.2788
PIM3	CUL4B	0.0019	0.2836	-0.2818
KLHL13	SKA3	-0.0729	0.2097	-0.2827
POGZ	CETN2	-0.1185	0.1715	-0.29
CCNDBP1	AVP11	0.0052	0.2975	-0.2923
STRADA	GNAI3	-0.1641	0.1631	-0.3271
MAPK6	CHAF1B	-0.2119	0.124	-0.3359
STAG3	HAUS7	-0.0422	0.3023	-0.3445
HAUS7	ANAPC2	-0.0151	0.3333	-0.3484
WDR6	RAD17	-0.22	0.1302	-0.3502
ENSA	ANAPC4	0.0228	0.3732	-0.3504
APITD1	STAG3	0.0103	0.3711	-0.3608
STAG3	CINP	-0.0812	0.4194	-0.5005

**Table 1.** Top 25 gene pairs whose  $\rho$  value differs between G1 and G2M stages.

Next we compare  $\rho$  values between MS participants and unaffected controls at the genus level. Table S7 shows the highest  $\rho$  values (in absolute value) in both MS and unaffected controls. The top 25 differences between MS and controls are shown in Table 2; a larger table with more genus pairs is shown in Table S8. Genera appearing among the top 25 differences that were previously shown to be associated with MS include *Megasphaera* [12]; *Acidaminococcus* [13]; *Ruminococcus* [3];



Dialister, Lachnospira, and Adlercreutzia [28]; Bacteroides and Prevotella [21, 28]; Lactobacillus [5], Adlercreutzia [28, 5]; and Eubacterium [23].

In all of the top 25 genus pairs, the estimated  $\rho$ -values are greater in controls compared to MS. The values for controls are positive and the MS values are either close to zero or negative. This trend is present in the top 100 genus pairs as well, with a few exceptions. In Figure S9, we see that there are only 25 genus pairs common in the top 100 pairs for MS and unaffected controls. This analysis has uncovered a strong disruption in community-level dynamics in the MS participants among a number of aforementioned genera whose abundances were already known to be related to MS.

Genus 1	Genus 2	$\rho_{MS}$	$\rho_C$	$\rho_{MS} - \rho_C$
UBA1417	BacteroidesB	-0.2295	0.2525	-0.4819
EubacteriumR	Prevotella	-0.203	0.2838	-0.4868
Dialister	UBA1417	-0.071	0.4202	-0.4913
UBA11774	Acidaminococcus	-0.0573	0.4377	-0.495
RuminiclostridiumE	LactobacillusB	-0.1886	0.3084	-0.4971
Senegalimassilia	Adlercreutzia	-0.2275	0.2735	-0.501
RuminiclostridiumC	CAG-83	0.0213	0.5311	-0.5098
PeH17	BacteroidesB	-0.2355	0.278	-0.5135
Adlercreutzia	Dialister	-0.293	0.2326	-0.5256
RuminiclostridiumE	Acidaminococcus	-0.0328	0.4996	-0.5323
Acidaminococcus	CAG-177	-0.0945	0.4461	-0.5406
Megasphaera	RuminococcusC	-0.0671	0.4807	-0.5478
Megasphaera	CAG-127	-0.0442	0.5057	-0.5499
Acidaminococcus	EubacteriumR	-0.2135	0.3547	-0.5683
Megasphaera	CAG-180	-0.1967	0.3853	-0.5819
Acidaminococcus	Adlercreutzia	-0.1906	0.413	-0.6036
EubacteriumR	Dialister	-0.2587	0.3482	-0.6069
Megasphaera	CAG-177	-0.1881	0.4253	-0.6135
Megasphaera	RuminococcusD	-0.1457	0.4686	-0.6143
Acidaminococcus	CAG-180	-0.2094	0.4055	-0.6149
RuminiclostridiumE	Megasphaera	-0.0811	0.5342	-0.6154
UBA11774	Lachnospira	-0.1046	0.5132	-0.6178
Acidaminococcus	RuminococcusC	-0.0643	0.563	-0.6273
Acidaminococcus	CAG-127	-0.0385	0.6221	-0.6606
UBA11774	Dialister	-0.1607	0.5012	-0.6619

**Table 2.** Top 25 genus pairs whose  $\rho$  value differs between MS ( $\rho_{MS}$ ) and control samples ( $\rho_C$ ).

## Conclusion

We have demonstrated that empirical estimates of metrics of proportionality in count-based platforms can lead to bias. Though the bias is mitigated when the mean read depths is large, there may still exist a bias in less-abundant features regardless. We therefore designed a multinomial logit-normal model to calculate model-based estimates of the proportionality metrics. We showed that, in an extensive simulation study, the model-based estimates outperformed the empirical estimates, even in the case of model misspecification. Additionally, a simple plugin estimator outperformed empirical estimates, which could be useful in the case where a user does not have the computational resources to obtain model estimates through the maximum likelihood estimator. Importantly, empirical vs. model-based estimates differed in a way that could drastically change results regarding which features are most strongly correlated with one another.

One limitation of our approach is that the model does not differentiate between structural and sampling zeros, hence the need for filtering out features with many zero counts. Though this could be remedied in the model by introducing zero-inflated parameters as done in [31, 16], this introduces a new problem regarding the interpretation of the proportionality metrics, which are not defined for zero-valued arguments. One could conceivably define these metrics conditional on non-zero values, but this would ignore valid information contained in the probability of a zero proportion. To address this, zero-inflated versions of Kendall's  $\tau$  and Spearman's  $\rho$  could be used in the context of a zero-inflated model. This is left for future work.

## Competing interests

H.T. has, in the last five years, received research support from the Canada Research Chair Program, the National Multiple Sclerosis Society, the Canadian Institutes of Health Research, the Multiple Sclerosis Society of Canada, the Multiple Sclerosis Scientific Research Foundation and the EDMUS Foundation ('Fondation EDMUS contre la sclérose en plaques'). In addition, in the last five years, has had travel expenses or registration fees prepaid or reimbursed to present at CME conferences from the Consortium of MS Centres (2018, 2023), the Canadian Neurological Sciences Federation (2023), National MS Society (2018, 2022), ECTRIMS/ ACTRIMS (2017-2023), American Academy of Neurology (2019). Speaker honoraria are either declined or donated to an MS charity or to an unrestricted grant for use by H.T.'s research group.

## Acknowledgments

K.M. acknowledges support from the Natural Sciences and Engineering Research Council of Canada (NSERC Discovery Grant: RGPIN-2021-03634) as well as an NSERC Undergraduate Student Research Award for N.O. The MS study was supported by funding from Canada Research Chair (PI: H.A.), the Multiple Sclerosis Society (PI: H.A.), and the Multiple Sclerosis Scientific and Research Foundation (#EGID: 2636; PI: H.T.). These latter funding sources were not involved in the study design, the collection, analysis, and interpretation of the data, or in the decision to submit this article for publication. We are grateful for all the participants' involvement, especially children and teenagers with MS and their parents. We are also grateful to all the investigators and study teams at each site involved in the Canadian Paediatric Demyelinating Disease Network study, without whom this study would not have been possible. We acknowledge the important contributions of the H.A. team (University of Manitoba) and H.T. team (University of British Columbia); Thomas Duggan in facilitating study set-up, coordination and data collection; Bonnie Leung for study coordination; Michael Sargent (Department of Internal Medicine, and the University of Manitoba IBD Clinical and Research Centre laboratory) for managing the stool biobank, and Jessica D. Forbes (University of Toronto) for assisting with the original grant.

## Data availability statement

The mouse embryonic stem cell dataset is publicly available under accession number E-MTAB-2512. The pediatric-onset MS dataset cannot be shared publicly for privacy of participants. The authors can be contacted for data access; requests will be assessed on a case-by-case basis, based on the scientific rigor of the research question.

## References

1. John Aitchison. *The statistical analysis of compositional data*. The Blackburn Press, 1986.
2. Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology*, 33(2):155–160, 2015.
3. Brandi L Cantarel, Emmanuelle Waubant, Christel Chehoud, Justin Kuczynski, Todd Z DeSantis, Janet

- Warrington, Arun Venkatesan, Claire M Fraser, and Ellen M Mowry. Gut microbiota in multiple sclerosis: possible influence of immunomodulators. *Journal of Investigative Medicine*, 63(5):729–734, 2015.
4. Jiahua Chen and Zehua Chen. Extended BIC for small-n-large-p sparse GLM. *Statistica Sinica*, pages 555–574, 2012.
5. Jun Chen, Nicholas Chia, Krishna R Kalari, Janet Z Yao, Martina Novotna, M Mateo Paz Soldan, David H Luckey, Eric V Marietta, Patricio R Jeraldo, Xianfeng Chen, et al. Multiple sclerosis patients have a distinct gut microbiota compared to healthy controls. *Scientific reports*, 6(1):1–10, 2016.
6. Juan José Egozcue, Jan Graffelman, M Isabel Ortego, and Vera Pawlowsky-Glahn. Some thoughts on counts in sequencing studies. *NAR Genomics and Bioinformatics*, 2(4):lqaa094, 2020.
7. Ionas Erb and Cedric Notredame. How should we measure proportionality on relative gene expression data? *Theory in Biosciences*, 135(1):21–36, 2016.
8. Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.
9. Gregory B Gloor, Jean M Macklaim, Vera Pawlowsky-Glahn, and Juan J Egozcue. Microbiome datasets are compositional: and this is not optional. *Frontiers in microbiology*, 8:2224, 2017.
10. Gustaf Hendeby and Fredrik Gustafsson. On nonlinear transformations of Gaussian distributions. *Technical Report from Automatic Control at Linköping University*, 2007.
11. Peter D Hoff. Nonparametric modeling of hierarchically exchangeable data. *University of Washington Statistics Department, Tech. Rep*, 421, 2003.
12. Sushrut Jangi, Roopali Gandhi, Laura M Cox, Ning Li, Felipe Von Glehn, Raymond Yan, Bonny Patel, Maria Antonietta Mazzola, Shirong Liu, Bonnie L Glanz, et al. Alterations of the human gut microbiome in multiple sclerosis. *Nature communications*, 7(1):12015, 2016.
13. Madina Kozhieva, Natalia Naumova, Tatiana Alikina, Alexey Boyko, Valentin Vlassov, and Marsel R Kabilov. Primary progressive multiple sclerosis in a Russian cohort: Relationship with gut bacterial diversity. *BMC microbiology*, 19(1):1–9, 2019.
14. David R Lovell, Xin-Yi Chua, and Annette McGrath. Counts: an outstanding challenge for log-ratio analysis of compositional data in the molecular biosciences. *NAR genomics and bioinformatics*, 2(2):lqaa040, 2020.
15. Aaron Lun. Overcoming systematic errors caused by log-transformation of normalized single-cell RNA sequencing data. *BioRxiv*, page 404962, 2018.
16. Siyuan Ma, Boyu Ren, Himel Mallick, Yo Sup Moon, Emma Schwager, Sagun Maharjan, Timothy L Tickle, Yiren Lu, Rachel N Carmody, Eric A Franzosa, et al. A statistical model for describing and simulating microbial community profiles. *PLOS Computational Biology*, 17(9):e1008913, 2021.
17. Kevin McGregor, Aurélie Labbe, and Celia MT Greenwood. MDiNE: a model to estimate differential co-occurrence networks in microbiome studies. *Bioinformatics*, 36(6):1840–1847, 2020.
18. Paul J McMurdie and Susan Holmes. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS computational biology*, 10(4):e1003531, 2014.
19. Ali I Mirza, Feng Zhu, Natalie Knox, Jessica D Forbes, Christine Bonner, Gary Van Domselaar, Charles N Bernstein, Morag Graham, Ruth Ann Marrie, Janace Hart, et al. The metabolic potential of the paediatric-onset multiple sclerosis gut microbiome. *Multiple Sclerosis and Related Disorders*, 63:103829, 2022.
20. Ali I Mirza, Feng Zhu, Natalie Knox, Jessica D Forbes, Gary Van Domselaar, Charles N Bernstein, Morag Graham, Ruth Ann Marrie, Janace Hart, E Ann Yeh, et al. Metagenomic analysis of the pediatric-onset multiple sclerosis gut microbiome. *Neurology*, 98(10):e1050–e1063, 2022.
21. Sachiko Miyake, Sangwan Kim, Wataru Suda, Kenshiro Oshima, Masakazu Nakamura, Takako Matsuoka, Norio Chihara, Atsuko Tomita, Wakiro Sato, Seok-Won Kim, et al. Dysbiosis in the gut microbiota of patients with multiple sclerosis, with a striking depletion of species belonging to clostridia XIVa and IV clusters. *PloS one*, 10(9):e0137429, 2015.
22. Thomas P Quinn, Ionas Erb, Greg Gloor, Cedric Notredame, Mark F Richardson, and Tamsyn M Crowley. A field guide for the compositional analysis of any-omics data. *GigaScience*, 8(9):giz107, 2019.
23. Marina Saresella, Ivana Marventano, Monica Barone, Francesca La Rosa, Federica Piancone, Laura Mendozzi, Alessia d’Arma, Valentina Rossi, Luigi Pugnetti, Gabriella Roda, et al. Alterations in circulating fatty acid are associated with gut microbiota dysbiosis and inflammation in multiple sclerosis. *Frontiers in immunology*, 11:1390, 2020.
24. Justin D Silverman, Kimberly Roche, Zachary C Holmes, Lawrence A David, and Sayan Mukherjee. Bayesian multinomial logistic normal models through marginally latent matrix-T processes. *J. Mach. Learn. Res.*, 23:7–1, 2022.
25. Michael A Skinnider, Jordan W Squair, and Leonard J Foster. Evaluating measures of association for single-cell transcriptomics. *Nature methods*, 16(5):381–386, 2019.
26. Gil Stelzer, Naomi Rosen, Inbar Plaschkes, Shahar Zimmerman, Michal Twik, Simon Fishilevich, Tsippi Iny Stein, Ron Nudel, Iris Lieder, Yaron Mazor, et al. The GeneCards suite: from gene data mining to disease genome sequence analyses. *Current protocols in bioinformatics*, 54(1):1–30, 2016.
27. Chuan Tian. Microbial network recovery by compositional graphical lasso under additive log-ratio transformation. 2020.
28. RE Ventura, T Iizumi, T Battaglia, Menghan Liu, GI Perez-Perez, J Herbert, and MJ Blaser. Gut microbiome of treatment-naïve MS patients of different ethnicities early in disease course. *Scientific reports*, 9(1):16396, 2019.
29. Amy D Willis and Bryan D Martin. Divnet: Estimating diversity in networked communities. *BioRxiv*, page 305045, 2018.
30. Fan Xia, Jun Chen, Wing Kam Fung, and Hongzhe Li. A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*, 69(4):1053–1063, 2013.
31. Yanyan Zeng, Daolin Pang, Hongyu Zhao, and Tao Wang. A zero-inflated logistic normal multinomial model for extracting microbial compositions. *Journal of the American Statistical Association*, pages 1–14, 2022.