Compositional analysis of microbiome data using the linear decomposition model (LDM)

Yi-Juan Hu\textsuperscript{1}\textsuperscript{*} and Glen A. Satten\textsuperscript{2}

\textsuperscript{1}Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, 30322, USA. Email: yijuan.hu@emory.edu.

\textsuperscript{2}Department of Gynecology and Obstetrics, Emory University School of Medicine, Atlanta, GA, 30322, USA. Email: gsatten@emory.edu.

\textsuperscript{*}Corresponding author

Keywords: compositional effect, log ratio, differential abundance, false discovery rate, community-level test
Abstract

Summary: There are compelling reasons to test compositional hypotheses about microbiome data. We present here LDM-clr, an extension of our linear decomposition model (LDM) approach to allow fitting linear models to centered-log-ratio-transformed taxa count data. As LDM-clr is implemented within the existing LDM program, it enjoys all the features supported by LDM, including a compositional analysis of differential abundance at both the taxon and community levels, while allowing for a wide range of covariates and study designs for either association or mediation analysis.

Availability and Implementation: LDM-clr has been added to the R package LDM, which is available on GitHub at https://github.com/yijuanhu/LDM.

Contact: yijuan.hu@emory.edu

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

Microbiome association studies detect taxa that are most strongly associated with the trait of interest (e.g., a clinical outcome or environmental factor), which are then used as microbial biomarkers for prognosis and diagnosis of a disease or microbial targets (e.g., pathogenic or probiotic bacteria) for drug discovery. Data derived from 16S amplicon or metagenomic sequencing are generally summarized in a taxa count table. However, since the total sample read count (library size) is an experimental artifact, only the relative abundances of taxa can be measured, not their absolute abundances. Consequently, microbiome data are compositional, meaning that a change in one taxon’s relative abundance necessitates a counterbalancing change in the relative abundance of (at least one) other taxa. The compositional null hypothesis tested for a taxon posits that the ratios of taxon relative abundances are constant across varying trait values. This hypothesis is frequently expressed in terms of parameters describing log-transformed data; the centered log ratio (clr), which centers the log relative abundance of each taxon by the sample mean log relative abundance of all taxa is most frequently used [1]. Since zero count data cannot be log-transformed, it is customary to add a pseudocount, usually 0.5 or 1, to the zeros or all entries in the taxa count table [1, 2].

We previously developed the linear decomposition model (LDM) [3–8] for testing hypotheses about the microbiome. Initially, LDM was designed for testing the null hypothesis that the relative abundance at a given taxon remains constant as the trait changes; this hypothesis could also be tested using arcsin-root-transformed relative abundance data. Later, LDM was extended to accommodate presence-absence data [5]. LDM is based on a linear model that regresses taxon data on traits/covariates, using a permutation procedure for inference. The linear model and permutation-based inference render LDM highly versatile, making it suitable for a wide variety of analyses (e.g., association, mediation), taxon data scales (e.g., relative abundance or presence-absence), trait types (e.g., continuous, discrete, multivariate, and time-to-event), and sample structures (e.g., matched sets, clustered data). Further, LDM
gives a unified taxon- and community-level analysis. A natural extension of LDM for compositional analysis would be to incorporate clr-transformed taxon data. Because the parameters tested in the compositional hypothesis involve more than one taxon, some changes to LDM are required, as described herein. The resulting tests still inherit the aforementioned features of LDM, as also summarized in Figure S1. In this paper, we present LDM-clr, an extension of LDM that allows analysis of clr data.

LinDA [9], a recently published method for compositional analysis of microbiome data, also relies on a linear model for clr-transformed data. It has demonstrated better FDR control and higher sensitivity than many existing compositional methods, including ANCOM-BC [2], ALDEx2 [10], metagenomeSeq [11], and MaAsLin2 [12]. However, LinDA bases its inference on asymptotic distributions, which may not be appropriate for small sample sizes, small numbers of taxa, and complex data structures. Furthermore, LinDA is not applicable for multivariate traits or time-to-event (survival) outcomes and cannot be used for testing community-level hypotheses.

Methods

Let \( Y_{ij} \) denote the number of sequence reads from sample \( i \) assigned to taxon \( j \) (\( j = 1, 2, \ldots, J \)), \( X_i \) represent a (possibly multivariate) trait of interest we wish to test, and \( Z_i \) indicate a set of potentially confounding covariates we wish to control for. Optimally, we would like to test hypotheses about how the absolute abundance of each taxon depends on the covariates. However, the information on absolute abundance is typically not available; even if the total bacterial counts are measured separately (e.g., using flow cytometry), the biases inherent in microbiome data make these values unreliable [13]. If instead we formulate a log-linear model for the observed counts \( Y_{ij} \), we must write

\[
\log(Y_{ij}) = \gamma_i + \alpha_j + X_i^T \beta_j + Z_i^T \beta_{Z,j} + \epsilon_{ij},
\]

where \( \gamma_i \) is an unknown, sample-specific size factor that encodes information on the ratio of observed to true counts, \( \beta_j \) is the association parameter of interest, \( \alpha_j \) and \( \beta_{Z,j} \) are nuisance
parameters and $\epsilon_{ij}$ is a mean-zero error. The intercept $\alpha_j$ is the sum of a term corresponding to the log relative abundance of taxon $j$ when $Z_i$ and $X_i$ are both zero and an unknown bias factor due to experimental bias [13–15]. The standard LDM cannot be used to fit (1) because of the $\gamma_i$s.

The compositional approach considers the $\gamma_i$s to be unestimable, and instead eliminates the $\gamma_i$s by writing a linear model for some form of log-ratio data. Here we consider the clr, given by

$$C_{ij} = \log(Y_{ij}) - J^{-1} \sum_{j'=1}^{J} \log(Y_{ij'})$$

Following from (1), we have

$$C_{ij} = \tilde{\alpha}_j + X_i^T \tilde{\beta}_j + Z_i^T \tilde{\beta}_{Z,j} + \tilde{\epsilon}_{ij},$$  

(2)

\[\tilde{\alpha}_j = \alpha_j - \bar{\alpha}, \quad \tilde{\beta}_j = \beta_j - \bar{\beta}, \quad \tilde{\beta}_{Z,j} = \beta_{Z,j} - \bar{\beta}_Z, \quad \text{and} \quad \tilde{\epsilon}_{ij} = \epsilon_{ij} - \bar{\epsilon}, \] where $\bar{\alpha}$, $\bar{\beta}$, $\bar{\beta}_Z$, and $\bar{\epsilon}$ are the averages of $\alpha_j$, $\beta_j$, $\beta_{Z,j}$, and $\epsilon_{ij}$ over all taxa, respectively. Model (2) can be fit using LDM, but the parameters are not the same as those in (1) as they are centered by their means over all taxa, which makes inference on individual taxa difficult. In particular, the usual procedure in LDM of testing $\tilde{\beta}_j = 0$ in (2) will not correspond to an easily interpretable result about the association between covariates $X_i$ and taxon $j$.

Denote the least-square estimators of $\tilde{\beta}_j$ and $\tilde{\beta}_{Z,j}$ obtained using the LDM by $\hat{\beta}_j$ and $\hat{\beta}_{Z,j}$, respectively. In order to obtain inference on individual taxa, we follow the approach in LOCOM [15] and assume that the majority (i.e., more than half) of the taxa are null taxa. This assumption has been frequently adopted in compositional methods (e.g., [15–17]). With this assumption, we expect that the median of the $\hat{\beta}_j$s, denoted by $\hat{\beta}^{(m)}$, will correspond to a taxon that is null. If this is the case, then we expect $\hat{\beta}_j - \hat{\beta}^{(m)} \approx 0$ if taxon $j$ is also null. When $\hat{\beta}_j$ is a vector, the median is calculated in an element-wise manner.

If the number of taxa $J$ is large, we can treat $\hat{\beta}^{(m)}$ as fixed; if we subtract $X_i^T \hat{\beta}^{(m)}$ from both sides of (2) then the estimated coefficients are automatically $\hat{\beta}_j - \hat{\beta}^{(m)}$. Then, LDM-clr performs permutation tests based on the proportion of permutation replicates in which the $F$ statistic

$$F_j = \frac{|\hat{\beta}_j - \hat{\beta}^{(m)}|^2}{\sum_i |C_{ij}|^2 - |\hat{\beta}_{Z,j}|^2 - |\hat{\beta}_j - \hat{\beta}^{(m)}|^2}$$  

(3)
exceeds the observed value, where \( \tilde{C}_{ij} = C_{ij} - X_i^T \tilde{\beta}^{(m)} \).

When the number of taxa \( J \) is small, \( \tilde{\beta}^{(m)} \) may contribute considerable variability. In this case, we still use (3), but use the value of \( \tilde{\beta}^{(m)} \) and hence \( \tilde{C}_{ij} \) calculated for each permutation replicate. Thus, the total sum-of-squares in the denominator of (3) is different for each replicate; the “missing” variability corresponds to the variability in \( \tilde{\beta}^{(m)} \). Similarly, the global test in LDM-clr sums numerators and denominators in (3) separately. The change in total variability \( \sum_i |\tilde{C}_{ij}|^2 \) across replicates does not affect the validity of the test because it is permutation-based. The flag \texttt{adjust.bias=TRUE} tells LDM-clr to use \( \tilde{\beta}^{(m)} \) from each replicate in the \( F \) statistic. In our simulations, this adjustment was only necessary for \( J < 100 \).

Like LOCOM and LDM-clr, LinDA [9] also bases inference on \( \tilde{\beta}_j - \tilde{\beta}^{(m)} \) although LinDA refers to \( \tilde{\beta} \) as a “bias” rather than recognizing that the regression coefficients in models (1) and (2) are simply different. LinDA estimates the “bias” \( \tilde{\beta} \) using the mode of the \( \tilde{\beta}_j \)'s, leading to the odd conclusion (in their framework) that the mode be used to estimate the mean of the \( \beta_j \) values. Still, in our framework, using the mode in place of the median when calculating \( \tilde{\beta}^{(m)} \) is attractive in some settings, and is offered as an option in LDM-clr.

\textbf{Results}

\textbf{Simulation studies}

We evaluated LDM-clr using the same set of simulation settings used to assess LinDA [9]: S0C0 (a binary trait), S0C1 (a continuous trait), S0C2 (a binary trait with two confounders), S1C0 (zero-inflated absolute abundances), S2C0 (correlated absolute abundances), S3C0 (Gamma abundance distribution), S4C0 (small number of taxa), S5C0 (small sample size), S6C0 (10-fold difference in library size), S7C0 (Negative-Binomial abundance distribution), S8.1C0 (pre-treatment and post-treatment comparison), and S8.2C0 (replicate sampling). We applied LDM-clr using both the mode and median as \( \tilde{\beta}^{(m)} \), referred to as LDM-mode and LDM-median, respectively, for testing individual taxa and the entire community. Additionally, we applied LinDA for testing individual taxa only.
First, we successfully replicated the results of LinDA as reported in [9]. In S4C0 (Figure 1, left), where LinDA yielded inflated FDR due to a small number of taxa, LDM-mode and LDM-median controlled the FDR as they adjusted for the uncertainty in the value of $\hat{\beta}(m)$. In S8.2 (Figure 1, right), where replicated samples were present, LDM-mode and LDM-median yielded higher sensitivity than LinDA, with an increase of up to 9%, which is attributable to our permutation-based inference. In general, whenever LinDA controlled the FDR, LDM-mode and LDM-median also controlled the FDR. When LinDA had inflated FDR, such as in S4C0 (Figures 1, left), S3C0 (Figure S7), and S6C0 (Figure S9), LDM-mode and LDM-median controlled the FDR or had less inflated FDR. The only exception is in S3C0 (Figure S7), where LDM-median had a more inflated FDR than LinDA. With sparse signals, use of the median for $\hat{\beta}(m)$ exhibited better sensitivity than use of the mode. Conversely, with dense signals, the mode-based estimator outperformed the median-based estimator. However, in both cases, using the median led to better power for community-level tests than using the mode (Figure S12).

**Compositional analysis of the aGVHD data**

We analyzed the 16S rRNA sequencing data on acute graft-versus-host disease (aGVHD) [18], which originally consisted 2436 operational taxonomic units (OTUs) across 94 subjects. Following the guidelines in [9], we removed subjects with library sizes less than 1000 and excluded OTUs found in fewer than 10% subjects, resulting in a final dataset comprising 89 subjects and 304 OTUs for our analysis. We tested the association of the gut microbiome with two survival outcomes separately: the overall survival and the time to stage-III aGVHD, with censoring rates of 52.3% and 42.0%, respectively. Both analyses were adjusted for age and gender.

We previously established in [7] that censored survival times can be tested in association with microbiome data using LDM by treating the Martingale or deviance residual obtained from the Cox model as a continuous trait; the Cox model incorporates potentially confounding
Figure 1: Empirical FDR and sensitivity at the nominal FDR level of 0.05 (black dashed line) in the S4C0 setting (with a small number of taxa) and the S8.2C0 setting (with replicated sampling). All results here as well as in other figures were based on 1000 replicates of data.
Table 1. Results from compositional analysis of the aGVHD data

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Method</th>
<th>Number of detected OTUs</th>
<th>Community-level p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Martingale</td>
<td>Deviance</td>
</tr>
<tr>
<td>Time to aGVHD</td>
<td>LDM-median</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>LDM-mode</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>LinDA</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>Overall survival</td>
<td>LDM-median</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>LDM-mode</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>LinDA</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Note: The nominal FDR level is 5%. “NA” indicates that the test is not available.

variables (such as age and gender in this case) as covariates, while excluding the microbiome data. LDM also provides a combination test that integrates the results from analyzing the two residuals. LDM-clr inherits these features, enabling testing of clr-transformed data with censored survival times at both taxon and community levels. Using the same approach, we can also apply LinDA to test censored survival times by manually calculating the Martingale or Deviance residual for these data, and then treating it as a covariate. However, LinDA does not support combination tests or community-level tests.

The results of our analyses are summarized in Table 1. For the time to stage-III aGVHD, LDM-mode detected 13 associated OTUs at the nominal FDR level of 5%, while LinDA detected 16, using the Martingale residual. Both methods detected 0 associated OTU using the Deviance residual, demonstrating a significant variability in results when choosing which one to use \textit{a priori}. Fortunately, LDM-mode offers a combination test that detected 8 associated OTUs. The community-level p-values generated by both LDM-median and LDM-mode were all small, aligning with the detection of associated OTUs. For the overall survival, LDM-mode detected 3 OTUs, consistent with its small community-level p-value, while LinDA detected none.

Discussion

Recently, Zhou et al. [9] have proposed LinDA, a linear model for clr-transformed data that has some similarities with LDM-clr. One major difference between LinDA and the LDM family
is that LinDA uses asymptotic distributional assumptions, so that $p$-values are calculated using normal (or $t$ distributions in the case of small sample sizes) for independent samples; inference in LDM is permutation-based. Additionally, for correlated samples, LinDA uses asymptotic results for the linear mixed-effect model, which may fail if the distributional assumptions do not hold; by contrast, LDM-clr handles sample clustering by a distribution-free approach [4]. Further, inference in LinDA does not account for variability in $\hat{\beta}^{(m)}$, which we find is important when the number of taxa is small.

Another difference between LinDA and LDM is that LDM provides a unified framework of taxon-level and community-level tests, ensuring coherence between their results; LinDA does not support community-level tests. Specifically, the test statistic for the community-level test is a ratio, with numerator and denominator given by the sum of the taxon-specific $F$-statistic numerators and denominators, respectively. When the number of taxa $J$ is large enough that the variability in $\hat{\beta}^{(m)}$ can be ignored, then the LDM-clr global test statistic is equal to the PERMANOVA test statistic based on Aitchison’s distance (the Euclidean distance of clr-transformed data) applied to data $\tilde{C}_{ij}$ (defined just after (3)). Here we note that the Freedman-Lane permutation scheme used by the LDM and PERMANOVA-FL functions in the R package LDM actually gives slightly higher power than that attained by the implementation of PERMANOVA in the R package vegan. Even when the variability in $\hat{\beta}^{(m)}$ is accounted for, LDM-clr still produces test statistics that are similar to PERMANOVA based on $\tilde{C}_{ij}$. We note that comparison of LDM-clr and PERMANOVA using Aitchison’s distance is appropriate as this distance is considered the appropriate metric for community-level tests in compositional analysis of microbiome data [19]. Thus, we are confident that LDM-clr provides a useful addition to the growing list of methods for compositional analysis of complex microbiome data.

**Funding**

This research was supported by the National Institutes of Health award R01GM141074...
(Hu, Satten).

References


10. Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. Unify-


19. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. Frontiers in microbiology. 2017;8:2224.