Identification of Separable OTUs for Multinomial Classification in Compositional Data Analysis

R. Alberich[1,2,3†], N.A. Cruz[1,2,3*†], R. Fernández[1†],
I. García Mosquera[1,2,3†], A. Mir[1,2,3†], F. Roselló[1†]

[1*]Department of Math and Computer Science, Universitat de les Illes Balears, Crta. Valldemossa, km. 7,5, Palma, 07122, Illes Balears, Spain.
[2]IDISBA Research Institute, Crta. Valldemossa, km. 79, Palma, 07120, Illes Balears, Spain.
[3]Institut d'Investigació en Intel·ligéncia Artificial (IAIB), Universitat de les Illes Balears.

*Corresponding author(s). E-mail(s): nelson-alirio.cruz@uib.es;
Contributing authors: r.alberich@uib.es; r.fernandez@uib.es;
irene.garcia@uib.es; arnau.mir@uib.es; frossello@me.com;
†These authors contributed equally to this work.

Abstract

High-throughput sequencing has transformed microbiome research, but it also produces inherently compositional data that challenge standard statistical and machine learning methods. Identifying microbial taxa that discriminate between biological or clinical groups requires methods that both respect compositional constraints and provide rigorous statistical inference. In this work, we propose a multinomial classification framework for compositional microbiome data based on penalized log-ratio regression and pairwise separability screening. The method quantifies the discriminative ability of each OTU through the area under the receiver operating characteristic curve (AUC) for all pairwise log-ratios and aggregates these values into a global separability index S_k, yielding interpretable rankings of taxa together with confidence intervals.

We illustrate the approach by reanalyzing the Baxter colorectal adenoma dataset and comparing our results with Greenacre's ordination-based analysis using Correspondence Analysis and Canonical Correspondence Analysis. Our models consistently recover a core subset of taxa previously identified as discriminant, thereby corroborating Greenacre's main findings, while also revealing additional OTUs that become important once demographic covariates are taken into account. In particular, adjustment for age, gender, and diabetes medication

improves the precision of the separation index and highlights new, potentially relevant taxa, suggesting that part of the original signal may have been influenced by confounding.

Overall, the integration of log-ratio modeling, covariate adjustment, and uncertainty estimation provides a robust and interpretable framework for OTU selection in compositional microbiome data. The proposed method complements existing ordination-based approaches by adding a probabilistic and inferential perspective, strengthening the identification of biologically meaningful microbial signatures.

1 Introduction

High-throughput sequencing technologies have made it possible to characterize microbial communities in great detail, typically through counts of operational taxonomic units (OTUs) (Caporaso et al. 2011). The resulting data are inherently compositional: each sample is constrained to a fixed total, a property known as the "closure" constraint (Aitchison and Bacon-Shone (1984), Gloor et al. (2017)). As a consequence, microbiome datasets generated by high-throughput sequencing should be treated as compositional at every stage of the analysis (Gloor et al. (2017), Martín-Fernández et al. (2015a)). When standard multivariate methods are applied directly to raw or proportion-normalized counts, they can introduce spurious associations and lead to misleading inferences (Gloor et al. 2017). This issue is particularly problematic in supervised learning, where the goal is often to identify a subset of OTUs that best discriminates between multiple phenotypic or clinical groups.

The challenge of identifying discriminative features from compositional data is not unique to microbiome research. Compositional structures arise in many other fields, including geochemistry, metabolomics, ecology, single-cell transcriptomics, and economics, whenever measurements represent parts of a whole (Greenacre et al. 2021). In all these settings, statistical and machine learning methods that ignore the compositional constraint risk producing biased estimates and incorrect interpretations (Quinn et al. 2018). There is therefore a clear need for classification and feature-selection frameworks that explicitly account for compositionality, in order to support valid inference, interpretability, and reproducibility in high-dimensional applications.

In microbiome studies in particular, researchers are often interested not only in accurate prediction but also in interpretable variable selection identifying "separable" OTUs whose relative abundances differ systematically across groups and that can be combined into discriminative microbial signatures. Existing approaches typically rely on marginal testing, methods developed for binary outcomes, or procedures that do not fully respect the compositional structure of the data (Susin et al. 2020). Moreover, they rarely offer a coherent way to quantify uncertainty in the selected OTUs or to propagate OTU-level variability into global measures of separability in multinomial classification problems.

In this work, we introduce a classification framework specifically designed for compositional microbiome data, with the aim of identifying microbial taxa that most effectively discriminate among multiple biological or clinical groups. The framework starts with standard preprocessing steps to ensure data quality and comparability, including the removal of rare taxa, Bayesian-multiplicative imputation of zeros, and centered log-ratio transformation to properly handle compositional constraints (Martín-Fernández et al. (2015a); —citeGloor2017).

At its core, the methodology uses a regression-based model that relates microbial composition to group membership while accounting for the inherent interdependence among components. Extending this model to the multinomial setting allows us to work naturally with more than two groups. By expressing the model in terms of log-ratios between pairs of OTUs, we can directly examine how the balance between taxa contributes to group differentiation (Greenacre (2021), Hinton and Mucha (2022)).

Building on this formulation, we propose a pairwise log-ratio screening procedure that measures the discriminative power of each OTU pair. For every pair, we assess its ability to separate groups using the area under the receiver operating characteristic curve (AUC) as a measure of separability. These pairwise results are then aggregated into a global separability index that summarizes the overall contribution of each OTU across all pairs, providing a data-driven ranking of the most informative taxa.

Finally, we incorporate formal procedures to quantify the uncertainty associated with the separability estimates, combining analytical variance estimators with resampling-based inference. This allows us to ensure that the OTUs we identify are not only discriminative but also statistically robust. The complete framework is implemented in the `codabiocom` R package, which offers tools for covariate adjustment, uncertainty quantification, and parallel computation, making it suitable for large-scale compositional datasets.

2 Preliminaries

Microbiome datasets obtained through high-throughput sequencing consist of abundance counts for m operational taxonomic units (OTUs) across n samples. These data are compositional by nature: only relative abundances are meaningful, and the total abundance within each sample is constrained to a fixed sum. Throughout this section, we describe the preprocessing steps and fundamental definitions used in our framework for OTU classification and selection.

2.1 Data Preprocessing

To ensure robustness and comparability across samples, the abundance matrix $\mathbf{X}$ of dimension $n \times m$ undergoes a series of preprocessing steps:

Filtering of rare OTUs. OTUs with extremely low prevalence contribute little information and increase noise. Those with two or fewer non-zero counts are excluded from the analysis, i.e., OTU j is removed if $\sum_{l=1}^{n} x_{lj} \leq 2$.

Zero imputation. Because compositional data analysis (CoDA) relies on logarithmic transformations, zero counts must be replaced with small positive values.

We employ the Bayesian-multiplicative imputation proposed by Martin-Fernandez et al. (2015b), which preserves the sample composition while maintaining relative ratios among non-zero components.

Centered log-ratio (clr) transformation. After imputation, each sample $\mathbf{x}_{i\cdot}$ is transformed using the clr transformation:

$$\tilde{x}_{ij} = \log\left(\frac{x_{ij}}{\left(\prod_{l=1}^{m} x_{il}\right)^{1/m}}\right), \quad i = 1, \ldots, n, \ j = 1, \ldots, m.$$

This yields the transformed matrix $\tilde{\mathbf{X}}$, ensuring scale invariance and compatibility with Euclidean geometry.

The first step of the workflow (Step 1 in Figure 1) involves removing low-frequency OTUs, imputing zeros, and applying the centered log-ratio (clr) transformation.

2.2 Basic Definitions

For clarity, we introduce terminology used throughout the paper.

Differential OTUs. A subset $S = \{\text{OTU}_{i_1}, \ldots, \text{OTU}_{i_k}\}$ is said to be *differential* if it provides measurable separation between two or more biological groups, according to a specified separability criterion.

Separability measure. For a given subset S, we define a separability index $\text{Sep}(S)$ as

$$\text{Sep}(S) = \frac{h}{\sqrt{h_1^2 + h_2^2}},$$

where h denotes the height of the dendrogram at which two groups become separated and h_i is the within-group height for group i. This index quantifies how distinctly the two clusters differ.

A subset S_1 is considered superior to S_2 if $\text{Sep}(S_1) > \text{Sep}(S_2)$.

3 Binary classification framework

We first describe our methodology in the binary setting, which provides the basis for its extension to multinomial classification. Let $\mathbf{X}$ be the $n \times m$ abundance matrix, with rows corresponding to samples and columns to OTUs, and let Y denote the group indicator taking values in $\{0, 1\}$. Our aim is to identify those OTUs (or combinations thereof) that best discriminate between the two groups, while respecting the compositional nature of the data.

3.1 CoDA-based regression for microbiome analysis

We address the classification problem using a penalized regression model in which discriminative OTUs are those with nonzero regression coefficients. More specifically, we build on the microbiome signature based on log-ratio analysis proposed in Calle et al. (2023).

We introduce the binary response variable Y as

$$Y = \begin{cases} 1, & \text{if sample } \mathbf{x}_{i\cdot} \in g_1, \\ 0, & \text{if sample } \mathbf{x}_{i\cdot} \in g_2, \end{cases}$$

where each sample $\mathbf{x}_{i\cdot}$ corresponds to the i-th row of matrix $\mathbf{X}$. The objective is to model Y based on the information in $\mathbf{X}$ within a generalized linear model framework that incorporates the compositional structure of the data.

A natural starting point is the log-contrast model (Aitchison and Bacon-Shone 1984; Lu et al. 2019), defined as

$$g(E(Y)) = \alpha_0 + \sum_{j=1}^{m} \alpha_j \log(X_j),$$

subject to the zero-sum constraint $\sum_{j=1}^{m} \alpha_j = 0$. This constraint enforces the scale-invariance principle required in compositional data analysis (CoDA), ensuring that the model depends only on log-ratios among components rather than on their absolute scale.

Equivalently, the model can be rewritten as a "log-ratio model", i.e., a generalized linear model involving all possible pairwise log-ratios (Bates and Tibshirani 2019):

$$g(E(Y)) = \theta_0 + \sum_{1 \leq i < j \leq m} \theta_{ij} \log\left(\frac{X_i}{X_j}\right). \tag{1}$$

This representation is particularly convenient in our context, as it explicitly encodes contrasts between OTUs.

Variable selection is then performed by fitting a penalized regression model to (1). The regression coefficients θ_{ij} are estimated by minimizing a loss function $L(\theta)$ with an elastic-net penalty:

$$\tilde{\theta} = \arg\min_{\theta} \left\{ L(\theta) + \lambda_1 \|\theta\|_2^2 + \lambda_2 \|\theta\|_1 \right\}. \tag{2}$$

The penalty parameters λ_1 and λ_2 can be reparametrized as

$$\lambda_1 = \frac{\lambda(1-\alpha)}{2}, \quad \lambda_2 = \lambda\alpha,$$

where λ controls the overall strength of the penalization and α determines the balance between ℓ_1- and ℓ_2-type regularization (Friedman et al. 2010). In the case of a linear regression model, the loss function takes the form

$$L(\theta) = \|Y - \mathbf{M}\theta\|,$$

5

where $\mathbf{M}$ is the matrix of all pairwise log-ratios, of dimension $n \times \frac{m(m-1)}{2}$. Extensions to logistic regression and other generalized linear models are discussed in Friedman et al. (2010) and apply directly in our setting.

This CoDA-based regression framework provides a global, multivariate signature of the microbiome. In the next subsection, we complement it with a pairwise log-ratio screening strategy that yields an interpretable separability index and allows for a fine-grained assessment of the discriminative power of individual log-ratios.

Step 2 in Figure 1 formalizes the compositional structure through a log-contrast regression model under a zero-sum constraint.

3.2 Pairwise log-ratio screening and separability index

After preprocessing and model specification, the core idea of our proposed analysis is to quantify the discriminative power of each pairwise log-ratio between OTUs. For each pair (j, j') with $1 \leq j < j' \leq m$, we proceed as follows:

1. **Construction of pairwise log-ratios.** For each sample, we compute the pairwise log-ratio
$$Z_{ijj'} = \log\left(\frac{x_{ij}}{x_{ij'}}\right).$$

2. **Fitting of (multi)class regression models.** We fit a logistic or multinomial logistic regression model for the response Y (the group label) using $Z_{ijj'}$ as predictor, optionally adjusted for additional covariates $\mathbf{X}$.

   In the multiclass case, for $Y \in \{1, \ldots, C\}$, the model takes the form, for each class $c \in \{1, \ldots, C-1\}$,

$$P(Y_i = c) = \frac{\exp(\beta_{0c} + \beta_{1c} Z_{ijj'} + \mathbf{X}_i^\top \boldsymbol{\gamma}_c)}{1 + \sum_{k=1}^{C-1} \exp(\beta_{0k} + \beta_{1k} Z_{ijj'} + \mathbf{X}_i^\top \boldsymbol{\gamma}_k)},$$

   with baseline category $Y_i = C$ given by

$$P(Y_i = C) = \frac{1}{1 + \sum_{k=1}^{C-1} \exp(\beta_{0k} + \beta_{1k} Z_{ijj'} + \mathbf{X}_i^\top \boldsymbol{\gamma}_k)}.$$

   Parameter estimation is performed via maximum likelihood. In practice, this model can be fitted in R using the `multinom()` function from the `nnet` package (Venables and Ripley 2002), which implements an efficient iteratively reweighted least squares (IRLS) algorithm.

3. **Computation of pairwise AUCs.** Using the fitted model, we obtain predicted probabilities for each class and evaluate the Area Under the Curve (AUC) as a measure of separability:

   - For binary classification $(C = 2)$, we use the standard AUC,

$$AUC_{jj'} = P(\text{score}^+ > \text{score}^-).$$

6

- For multiclass classification ($C > 2$), we employ the generalized AUC proposed by Hand and Till (2001),

$$AUC_{jj'} = \frac{2}{C(C-1)} \sum_{c<c'} AUC(c, c'),$$

where $AUC(c, c')$ denotes the binary AUC comparing classes c and c'.

This procedure yields a symmetric $m \times m$ matrix $\mathbf{A} = [AUC_{jj'}]$, which summarizes the pairwise discriminative ability of OTU log-ratios. OTUs are then ranked according to the sum of each column of $\mathbf{A}$, prioritizing variables that consistently participate in high-scoring pairs.

Separability index.

To determine an optimal set of OTUs, we define a separability index as the average pairwise AUC among the top k ranked OTUs:

$$S_k = \frac{2}{k(k-1)} \sum_{1 \le j < j' \le k} AUC_{jj'}.$$

The final relevant set is chosen up to the value k at which S_k reaches its maximum, providing a data-driven criterion for the number of OTUs to retain.

Steps 3 and 4 (Figure 1) describe the pairwise log-ratio screening and ranking procedure based on the separability index S_k.

3.3 Uncertainty quantification

An important aspect of the proposed methodology is the estimation of the uncertainty associated with the pairwise AUCs and with the derived separability index S_k. Since the construction of S_k involves multiple levels of aggregation, we distinguish two main sources of dependence.

1. **Pairwise AUC variance in binary and multiclass settings.** For each pair of OTUs (j, j'), a pairwise log-ratio predictor is constructed and a classification model is fitted to predict Y. The resulting AUC, $AUC_{jj'}$, quantifies the ability of this log-ratio to discriminate between classes.

- **Binary case ($C = 2$).** Two common approaches exist for estimating $\mathrm{Var}(\widehat{AUC}_{jj'})$.
  First, the classical approximation of Hanley and McNeil (1982):

$$\mathrm{Var}_{\mathrm{Hanley}}(\widehat{AUC}_{jj'}) = \frac{AUC_{jj'}(1 - AUC_{jj'}) + (n_1 - 1)(Q_1 - AUC_{jj'}^2)}{n_1 n_0} + \frac{(n_0 - 1)(Q_2 - AUC_{jj'}^2)}{n_1 n_0}, \tag{3}$$

$$Q_1 = \frac{AUC_{jj'}}{2 - AUC_{jj'}}, \quad Q_2 = \frac{2\,AUC_{jj'}^2}{1 + AUC_{jj'}}.$$

Second, the nonparametric U-statistic approach of DeLong et al. (1988), for which a fast implementation is given by Sun and Xu (2014). Define

$$V_i = \frac{1}{n_0} \sum_{j=1}^{n_0} \mathbf{1}\{S_i^+ > S_j^-\}, \quad W_j = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{1}\{S_i^+ > S_j^-\},$$

where S_i^+ are the classifier scores on the n_1 positive cases and S_j^- on the n_0 negative cases. Then

$$\mathrm{Var}_{\mathrm{DeLong}}(\widehat{AUC}_{jj'}) = \frac{1}{n_1(n_1 - 1)} \sum_{i=1}^{n_1} (V_i - \widehat{AUC}_{jj'})^2 + \frac{1}{n_0(n_0 - 1)} \sum_{j=1}^{n_0} (W_j - \widehat{AUC}_{jj'})^2.$$
(4)

- **Multiclass case ($C > 2$).** When the response is multiclass, we use the Hand–Till generalization
$$AUC_{jj'} = \frac{2}{C(C - 1)} \sum_{c < c'} AUC_{cc'},$$

where $AUC_{cc'}$ is the binary AUC comparing classes c and c'. The variance of $AUC_{jj'}$ then propagates the uncertainty of each binary component:

$$\mathrm{Var}(AUC_{jj'}) = \frac{4}{[C(C - 1)]^2} \left[\sum_{c < c'} \mathrm{Var}(AUC_{cc'}) + 2 \sum_{\substack{(c < c') < (l < l') \\ (c,c') \neq (l,l')}} \mathrm{Cov}(AUC_{cc'}, AUC_{ll'}) \right].$$

Here, $\mathrm{Var}(AUC_{cc'})$ can be approximated using the binary formula in (3), while the covariances between binary pairs need to be approximated or estimated.

2. **Dependence among AUCs across OTU pairs.** Each $AUC_{jj'}$ shares OTUs with other pairs, inducing dependence in the matrix $\mathbf{A}$. When aggregating AUCs to compute the separability index S_k, this dependence propagates to its variance.

   Recall that
$$S_k = \frac{2}{k(k - 1)} \sum_{j < j' \leq k} AUC_{jj'},$$

so that

$$\mathrm{Var}(S_k) = \frac{4}{[k(k - 1)]^2} \left[\sum_{j < j'} \mathrm{Var}(AUC_{jj'}) + 2 \sum_{(j < j') < (l < l')} \mathrm{Cov}(AUC_{jj'}, AUC_{ll'}) \right].$$

Pairs that share at least one OTU (e.g., (AUC_{12}, AUC_{13})) are typically positively correlated. In practice, these covariances are non-negligible but difficult to derive

analytically. A feasible approximation is

$$\mathrm{Cov}(AUC_{jj'}, AUC_{ll'}) \approx \rho_{OTU}\sqrt{\mathrm{Var}(AUC_{jj'})\,\mathrm{Var}(AUC_{ll'})}$$

if the pairs share at least one OTU, where ρ_{OTU} can be estimated empirically from resampling runs or set to a conservative value (e.g., 0.1–0.3).

For the multiclass case, an additional layer of dependence arises because the Hand–Till AUC is itself an average of dependent binary pairs. An analogous approximation can be used for $\mathrm{Cov}(AUC_{cc'}, AUC_{ll'})$ when the class pairs share at least one class.

3. **Alternative: non-parametric bootstrap.** As an alternative to analytic variance formulas, we consider a non-parametric bootstrap procedure. This approach is particularly useful in capturing the sampling variability of S_k under complex dependency structures:

- Resample the n observations with replacement to generate bootstrap samples (optionally in a stratified manner by class).
- For each bootstrap replicate, recompute the full pairwise AUC matrix and the corresponding S_k.
- Estimate $\mathrm{Var}(S_k)$ using the empirical variance of the bootstrap replicates.

In the presence of class imbalance, a stratified bootstrap—resampling independently within each class—helps preserve marginal class frequencies and avoid inflated variance estimates.

Step 5 (Figure 1) addresses variance estimation and uncertainty quantification of both $AUC_{jj'}$ and S_k.

An additional advantage of the proposed strategy is that it naturally accommodates covariates in the multinomial regression model used to estimate class probabilities for each log-ratio pair. This allows one to adjust for potential confounding variables (such as age, sex, or clinical covariates) when assessing the discriminative ability of the OTU pairs. In the implementation provided by the `codabiocom` package, the `rowlogratios()` function accepts an optional covariate matrix $\mathbf{X}$, which is included alongside each pairwise log-ratio in the multinomial model. As a result, the computed AUC values reflect the marginal contribution of each log-ratio pair beyond the effect of included covariates, within a workflow that is fully compatible with parallel computation.

To summarize the proposed methodology and clarify its main components, Figure 1 presents an overview of the complete classification framework for microbiome data. Each step corresponds to one of the methodological sections described in detail below, from preprocessing and log-ratio modeling to uncertainty quantification and computational implementation.

Objective: Identify differential OTUs that best discriminate between groups $(g = 1, \ldots, k)$ using CoDA and pairwise log-ratio modeling.

Input data: Abundance matrix $\mathbf{X}_{n \times m}$ where rows = samples, columns = OTUs, and class labels $Y \in \{1, \ldots, k\}$.

1. Data Preprocessing:
(a) Remove OTUs with $\sum_i x_{ij} \leq 2$;
(b) Replace zeros using Bayesian-multiplicative imputation (Martin-Fernandez et al. 2015b);
(c) Apply centered log-ratio (clr) transformation.

2. Log-Contrast Regression Model:
Model the categorical response Y via a penalized log-contrast regression:

$$g(E[Y]) = \alpha_0 + \sum_{i=1}^{m} \alpha_i \log(X_i), \quad \text{s.t.} \sum_i \alpha_i = 0.$$

For $k > 2$, use multinomial logistic link $g(\cdot)$. *Equivalent form:* pairwise log-ratio model

$$g(E[Y]) = \theta_0 + \sum_{i<j} \theta_{ij} \log\left(\frac{X_i}{X_j}\right),$$

estimated using elastic-net penalization (Friedman et al. 2010).

3. Pairwise Log-Ratio Screening:
For each OTU pair (j, j'):

1. Compute log-ratio $Z_{ijj'} = \log(x_{ij}/x_{ij'})$;
2. Fit logistic or multinomial model $Y \sim Z_{ijj'} + \mathbf{X}$;
3. Compute the AUC ($AUC_{jj'}$) for class discrimination.

Obtain the symmetric AUC matrix $\mathbf{A} = [AUC_{jj'}]$.

4. OTU Ranking and Selection:
Rank OTUs by total AUC contribution. Define separability index:

$$S_k = \frac{2}{k(k-1)} \sum_{1 \leq j < j' \leq k} AUC_{jj'},$$

and select the number k that maximizes S_k.

5. Variance and Uncertainty Estimation:
Estimate $\text{Var}(AUC_{jj'})$ using:

- Hanley and McNeil (1982) DeLong et al. (1988) formulas;
- Propagation to $\text{Var}(S_k)$ including OTU-level covariance;
- Alternatively, stratified bootstrap resampling.

6. Output:
$\Rightarrow$ Optimal subset of differential OTUs maximizing separability S_k;
$\Rightarrow$ Statistical uncertainty quantified by analytical or bootstrap variance.

Software: Implemented in the codabjjpcom R package, supporting covariate adjustment and parallel processing.

Fig. 1 Overview of the proposed classification framework for microbiome data.

As outlined in Figure 1, the workflow proceeds in six main steps: data preprocessing, log-contrast modeling, pairwise log-ratio screening, ranking and selection via the separability index S_k, uncertainty quantification, and final output generation. The following sections detail each of these steps.

3.4 Computational complexity and parallel implementation

The pairwise log-ratio approach inherently involves a computational cost that grows quadratically with the number of OTUs, since the total number of unique pairs is $\binom{m}{2} = O(m^2)$. This can become a bottleneck when working with large, high-dimensional microbiome datasets.

To address this, the pairwise AUC computation has been implemented in a fully parallelized framework using the `foreach` and `doParallel` packages in R. The core operation is handled by the `rowlogratios()` function, which fits the (multinomial) regression model and computes the pairwise AUC for each log-ratio. The `calcAUClr()` function orchestrates the parallel computation across all OTU pairs, efficiently distributing tasks across multiple cores and thus exploiting available computational resources.

The final step (Step 6 in Figure 1) focuses on computational scalability and parallelization, ensuring the feasibility of the method for high-dimensional microbiome datasets.

This design enables not only the calculation of the initial AUC matrix, but also makes bootstrap resampling of the separability index S_k feasible at scale. By resampling the data and recomputing the entire AUC matrix in each replicate, this strategy naturally captures all levels of dependence: among OTU pairs, within multiclass combinations, and across the final index calculation.

All steps are implemented in the open-source `codabiocom` R package, which is freely available at

https://github.com/Cruzalirio/codabiocom

ensuring that the proposed methodology is fully reproducible and scalable to large compositional datasets.

4 Application to microbiome data

4.1 Data description

We illustrate the proposed methodology using the colorectal cancer screening dataset originally reported by Baxter et al. (2016) and later reanalyzed by Greenacre (2021). The data consist of operational taxonomic unit (OTU) abundances for 336 microbial features measured on stool samples from subjects classified as adenoma or control. In addition to the compositional microbiome data, the dataset includes several covariates, such as age, gender, and diabetes medication usage.

Our primary objective is to classify subjects into adenoma versus control using the compositional OTU data, while appropriately adjusting for potential confounders.

To this end, we applied the `LRRelev` procedure, which combines log-ratio logistic regression with the Hanley separation index. Four different modeling strategies were considered:

- **Model A**: unadjusted (no covariates),
- **Model B**: adjusted for age,
- **Model C**: adjusted for age and gender,
- **Model D**: adjusted for age, gender, and diabetes medication.

In all cases, the OTU data were transformed into pairwise log-ratios using the `rowlogratios()` function in `codabiocom`, and the corresponding separation indices were computed following the methodology described in the previous sections. All computations were performed using six processing cores to ensure adequate computational efficiency.

4.2 OTU relevance across models

Each model produces a relevance ranking of OTUs based on the separability index, with confidence intervals derived from the estimated variance. Table 1 summarizes, for each model, the number of relevant OTUs selected and the top ranked taxa.

Table 1 Number and top relevant OTUs (ordered by association index) per model.

Model	Relevant OTUs	Top OTUs (ordered)
A (Unadjusted)	7	Otu000105, Otu000310, Otu000281, Otu000264, Otu000058, Otu000113, Otu000067
B (Age only)	17	Otu000310, Otu000105, Otu000281, Otu000264, Otu000113, Otu000260, Otu000286, Otu000058, Otu000067, Otu000057, Otu000356, Otu000097, Otu000094, Otu000053, Otu000297, Otu000054, Otu000278
C (Age, Gender)	9	Otu000310, Otu000105, Otu000281, Otu000264, Otu000113, Otu000260, Otu000286, Otu000058, Otu000067
D (Age, Gender, Diabetes)	9	Otu000310, Otu000105, Otu000281, Otu000264, Otu000113, Otu000260, Otu000286, Otu000058, Otu000067

The consistency in the selection of Otu000310, Otu000105, Otu000281, Otu000264, Otu000113, Otu000067, and Otu000058 across all models highlights their robustness as discriminant features for diagnosis classification. The inclusion of covariates such as age, gender, and diabetes medication leads to changes in the number of selected OTUs and affects their ranking, suggesting that some taxa may be partially confounded by demographic or clinical factors.

Table 2 summarizes the estimated separation index and its confidence interval for each model.

Table 2 Separation index and 95% confidence intervals per model.

Model	Relevant OTUs	Separation index	Lower CI	Upper CI
A (Unadjusted)	7	0.5816	0.567	0.596
B (Age only)	17	0.6685	0.655	0.682
C (Age, Gender)	9	0.6962	0.683	0.709
D (Age, Gender, Diabetes)	9	0.7001	0.687	0.713

The separation index estimates and their associated confidence intervals provide insight into the robustness and precision of OTU-based discrimination between adenoma and control groups. Model A, which does not adjust for any covariates, shows moderate separation performance, with a separation index around 0.58. Adjustment for demographic covariates substantially improves discrimination: Model B, adjusting only for age, increases the separation index to approximately 0.67, while Models C and D (adjusting for age and gender, and for age, gender, and diabetes medication, respectively) yiel

5 Discussion

The findings of this study highlight several important aspects of the proposed framework for identifying discriminant OTUs in compositional microbiome data.

First, a core set of taxa consistently emerged across all models, even after adjusting for demographic and clinical covariates. This consistency indicates that these microbial features are robustly associated with disease status and are likely to represent biologically meaningful signals rather than artifacts of data preprocessing or model specification.

Second, the influence of covariates on model performance and feature selection is evident. Accounting for variables such as age, gender, and diabetes medication improves model precision reflected in narrower confidence intervals for the separability index S_k and reveals additional taxa that were not detected in unadjusted analyses. This result reinforces the importance of incorporating relevant covariates in microbiome studies to reduce confounding and to improve the interpretability of microbial associations with health outcomes.

Third, the proposed framework adds an element of statistical interpretability that complements existing ordination-based techniques such as Correspondence Analysis (CA) and Canonical Correspondence Analysis (CCA). The Hanley separation index provides an intuitive, non-parametric measure of group separation that directly quantifies discriminatory power, while the associated confidence intervals enable formal inference. This combination bridges exploratory visualization and rigorous statistical modeling, making the results both interpretable and statistically grounded.

From a biological standpoint, the persistence of certain taxa across multiple adjusted models suggests that these organisms may play a genuine role in disease development or progression. Their stability across analytical settings supports their

potential use as biomarkers in diagnostic or preventive microbiome research, and as targets for future mechanistic or interventional studies.

Taken together, these results demonstrate that integrating log-ratio modeling with covariate adjustment and uncertainty quantification provides a statistically principled and biologically meaningful approach for OTU selection in compositional data. The proposed methodology enhances both interpretability and reproducibility, which are essential for advancing microbiome-based inference and discovery.

6 Conclusions and future work

This study presents a unified framework for multinomial classification and feature selection in compositional microbiome data. By combining log-ratio transformations, penalized regression, and pairwise AUC-based separability measures, the approach offers a balance between computational feasibility, interpretability, and statistical rigor. The separability index S_k provides a simple yet effective way to rank and select discriminant taxa, while analytic and bootstrap-based uncertainty estimates ensure that the conclusions drawn are statistically robust.

Our application to the Baxter dataset confirms that the proposed approach not only reproduces known microbial patterns but also enhances inference by adjusting for covariates and providing confidence intervals. In this sense, the method complements ordination-based tools by offering a probabilistic perspective that directly connects microbial compositions to clinical outcomes.

Looking ahead, several extensions are worth pursuing. One direction involves extending the framework to hierarchical taxonomic levels (e.g., genus or family), which could improve biological interpretation. Another promising line is the incorporation of Bayesian priors to stabilize feature selection in high-dimensional contexts. Scalability to other omics domains such as metagenomics, metabolomics, or transcriptomics will also broaden the applicability of the method. Finally, integrating cross-study validation procedures could help assess the generalizability and reproducibility of the identified microbial signatures.

Overall, the proposed framework provides a transparent, reproducible, and statistically grounded approach for compositional data classification, implemented in the open-source `codabiocom` R package. It represents a step toward more interpretable and reproducible microbiome analytics, bridging methodological rigor with biological insight.

Declarations

References

Aitchison, J., Bacon-Shone, J.: Log contrast models for experiments with mixtures. Biometrika **71**(2), 323–330 (1984) https://doi.org/10.1093/biomet/71.2.323 https://academic.oup.com/biomet/article-pdf/71/2/323/635145/71-2-323.pdf

Baxter, N.T., Ruffin IV, M.T., Rogers, M.A., Schloss, P.D.: Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. Genome medicine **8**(1), 37 (2016)

Bates, S., Tibshirani, R.: Log-ratio lasso: Scalable, sparse estimation for log-ratio models. Biometrics (2), 613–624 (2019) https://doi.org/10.1111/biom.12995

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J., Fierer, N., Knight, R.: Global patterns of 16s rrna diversity at a depth of millions of sequences per sample. Proceedings of the National Academy of Sciences of the United States of America **108**(Supplement 1), 4516–4522 (2011) https://doi.org/10.1073/pnas.1000080107

Calle, M.L., Pujolassos, M., Susin, A.: coda4microbiome: compositional data analysis for microbiome cross-sectional and longitudinal studies. BMC Bioinformatics **24**(82) (2023) https://doi.org/10.1186/s12859-023-05205-3

DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, 837–845 (1988)

Friedman, J.H., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software **33**(1), 1–22 (2010) https://doi.org/10.18637/jss.v033.i01

Greenacre, M., Martínez-Álvaro, M., Blasco, A.: Compositional data analysis of microbiome and any-omics datasets: A validation of the additive logratio transformation. Frontiers in Microbiology **12**, 727398 (2021) https://doi.org/10.3389/fmicb.2021.727398

Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., Egozcue, J.J.: Microbiome datasets are compositional: And this is not optional. Frontiers in Microbiology **8**, 2224 (2017) https://doi.org/10.3389/fmicb.2017.02224

Greenacre, M.: Compositional data analysis. Annual Review of Statistics and its Application **8**(1), 271–299 (2021)

Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (roc) curve. Radiology **143**(1), 29–36 (1982)

Hinton, A.L., Mucha, P.J.: A simultaneous feature selection and compositional association test for detecting sparse associations in high-dimensional metagenomic

data. Frontiers in Microbiology **13**, 837396 (2022)

Hand, D.J., Till, R.J.: A simple generalisation of the area under the roc curve for multiple class classification problems. Machine learning **45**(2), 171–186 (2001)

Lu, J., Shi, P., Li, H.: Generalized linear models with linear constraints for microbiome compositional data. Biometrics **75**(1), 235–244 (2019) https://doi.org/10.1111/biom.12956 https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.12956

Martín-Fernández, J.A., Hron, K., Templ, M., Filzmoser, P., Palarea-Albaladejo, J.: Bayesian-multiplicative treatment of count zeros in compositional data sets. Statistical Modelling **15**(2), 134–158 (2015) https://doi.org/10.1177/1471082X14535524

Martin-Fernandez, J.-A., Hron, K., Templ, M., Filzmoser, P., Palarea-Albaladejo, J.: Bayesian-multiplicative treatment of count zeros in compositional data sets. Statistical Modelling **15**(2), 134–158 (2015) https://doi.org/10.1177/1471082X14535524 https://doi.org/10.1177/1471082X14535524

Quinn, T.P., Erb, I., Richardson, M.F., Crowley, T.M.: Understanding sequencing data as compositions: an outlook and review. Bioinformatics **34**(16), 2870–2878 (2018) https://doi.org/10.1093/bioinformatics/bty175

Susin, A., Wang, Y., Lê Cao, K.-A., Calle, M.L.: Variable selection in microbiome compositional data analysis. NAR Genomics and Bioinformatics **2**(2), 029 (2020) https://doi.org/10.1093/nargab/lqaa029

Sun, X., Xu, W.: Fast implementation of delong?s algorithm for comparing the areas under correlated receiver operating characteristic curves. IEEE Signal Processing Letters **21**(11), 1389–1393 (2014)

Venables, W.N., Ripley, B.D.: Modern Applied Statistics with S, 4th edn. Springer, New York (2002). ISBN 0-387-95457-0. https://www.stats.ox.ac.uk/pub/MASS4/