A TENSOR DECOMPOSITION MODEL FOR LONGITUDINAL MICROBIOME STUDIES

BY SIYUAN MA AND HONGZHE LI

Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania
siyuan.ma@pennmedicine.upenn.edu; hongzhe@pennmedicine.upenn.edu

Longitudinal microbiome studies can help delineate true biological signals from the high interindividual variability that is common in microbiome data. However, there are few methods available for unsupervised dimension reduction of time course microbial abundance observations. Existing methods do not fully observe the distribution characteristics of such data types, namely, zero-inflation, compositionality, and overdispersion. We present a tensor decomposition model and a semiparametric quasi-likelihood estimation method for the decomposition of longitudinal microbiome data, generalizing existing approaches in tensor decomposition of Gaussian data. Optimization is performed through projected gradient descent additionally allowing interpretability constraints. We show through simulation studies our method is able to recover low rank structures from microbiome time course data, better than existing approaches. Lastly, we apply our method to two existing longitudinal microbiome studies, to detect global microbial changes associated with dietary and pharmaceutical effects, as well as infant birth modes.

1. Introduction. Study of the human associated microbiota has greatly benefited from epidemiology cohorts (Cho and Blaser (2012)). With such cohorts, changes in the abundance of microbial organisms (“features”) in human environments, measured via high throughput sequencing (Kuczynski et al. (2012)), have been linked to a wide range of host health conditions (Kostic, Xavier and Gevers (2014); Giongo et al. (2011)) and treatment effects (Francino (2016)). In recent years, microbiome epidemiology has begun to expand from traditional cross-sectional comparisons to longitudinal follow-up designs (Lloyd-Price et al. (2019); Kostic et al. (2015); Tanes et al. (2021)). Compared to cross-sectional cohorts, longitudinally sampled microbiomes can model ecological development dynamics (Kostic et al. (2015)), as well as control for the large, confounding inter-subject variability that has been noted for such data (Tanes et al. (2021)). In practice, longitudinal designs have facilitated recent discoveries such as overall ecological dysbiosis in gut community structure during inflammatory bowel diseases (IBD) (Lloyd-Price et al. (2019)), and microbial modulations of nutritional protections against cardiometabolic disease (Wang et al. (2021)).

Methodology-wise, statistical research for longitudinal microbiomes have focused on supervised differential abundance testing, i.e., linking individual microbial features to host conditions (Chen and Li (2016); Zhang et al. (2018)). Unsupervised dimension reduction, one of the most common microbiome analysis tasks (McMurdie and Holmes (2013)), has however not been well studied in the longitudinal setting. Traditional dimension reduction is performed through either clustering (Holmes, Harris and Quince (2012)) or ordination analysis (Gower (1966)). Of the two, ordination is agnostic towards the existence or lack of microbial clustering, and provides low-rank, biologically interpretable representation of high-dimensional datasets. It has been adopted in applications such as data visualization (Consortium et al. (2012)), characterizing disease dysbiosis (Gevers et al. (2014)), and microbiome-related mediation analysis (Zhang, Wei and Chen (2018)). Ordination dimension reduction

Keywords and phrases: Metagenomics, Dimension reduction, Quasi-likelihood.
for longitudinal microbiomes require special considerations to accommodate within-subject correlations (Chen and Li (2016)). Methods should also address common issues of microbiome sequencing data types (Mallick et al. (2017)), including: 1) sparsity, or zero-inflation (microbial features often have zero count observations, due to either absence or extreme low abundance), 2) compositionality (per-sample feature abundances can only be measured on the relative scale with sequencing technology, and are constrained to sum up to one), and 3) overdispersion (sequencing counts often have high variabilities that are not well approximated by restrictive parametric models). For ordination dimension reduction of longitudinal microbiome studies, both the common data issues and design-specific within-subject correlations should be addressed.

To this end, few dedicated methods are available. Previously, the canonical polyadic decomposition (CPD, Hitchcock (1927)) has been proposed as a popular method for generic datasets with a tensor-like structure. In the context of longitudinal designs, such methods assume that the data, viewed as an order-three tensor (feature × subject × timepoint), can be decomposed into the sum of rank-1 factors, each expressed as the outer product of feature-, subject-, and time-specific loadings. Within-subject correlation is thus naturally accounted for, and dimension reduction can be achieved by approximating the original data with factors of the highest variability. The CPD has been adapted to molecular epidemiology applications such as gene expressions across multiple tissues of the same subject (Wang, Fischer and Song (2019)). Only recently, Martino et al. (2021) proposed the compositional tensor factorization (CTF), which paired CPD with a modified centered log ratio transformation to perform dimension reduction on microbial sequencing counts. Compared to off-the-shelf CPD approaches, CTF explicitly addresses the zero-inflated and compositional nature of microbiome datasets by modifying the normalizing factor in the centered log ratio transformation. However, as we will discuss in more details in Section 4, the new normalization factor introduces additional, unnecessary variability that causes to non-trivial efficiency loss. No special considerations were made to account for data overdispersion in this method, either.

Motivated by two microbiome longitudinal studies, namely, the FARMM cohort (Tanes et al. (2021) and the DIABIMMUNE cohort (Kostic et al. (2015), we propose a dimension reduction method specialized for longitudinal microbiome studies, named “microTensor” (tensor decomposition for microbiome data). Similar to existing molecular epidemiology applications (Wang, Fischer and Song (2019); Martino et al. (2021)), microTensor generalizes the CPD approach, but with a few important, microbiome-specific considerations. First, we assume a mean structure model that addresses the sparsity and compositionality of microbial count observations, while adopting the CPD decomposition to account for within-subject correlations. Our objective function also does not require intermediate estimation of per-sample normalization factors. Second, we adopt a semiparametric quasi-likelihood approach to accommodate additional variabilities in the data, further improving estimation efficiency. Last, we propose an accompanying projected gradient descent algorithm that accommodates the identifiability requirements of our model, while also allowing for additional interpretability constraints. Compared to existing approaches, microTensor is uniquely appropriate for correlated microbial observations, addressing the aforementioned data-specific characteristics and facilitating efficient, interpretable ordination dimension reduction.

2. The FARMM study, data and variables. We first describe the FARMM microbiome cohort, and use it to introduce data notations and motivate our dimension reduction method. The Food and Resulting Microbial Metabolites (FARMM) study (Tanes et al. (2021)) followed subjects under three different dietary regimes (10 subjects each): omnivore, vegan, and EEN (a liquid diet with no fiber). Daily stool microbial samples were obtained over the study period of 15 days. Additionally, antibiotics and polyethylene glycol were administered to all subjects during days 6, 7, and 8, to transiently reduce bacterial load within the
gut, thus allowing for the assessment of microbiota recovery during days 9 to 15. Microbes were measured with metagenomic shotgun sequencing, which provided species level feature abundances. For filtering, we removed features present in no more than five samples ($\sim 1\%$) with at least $10^{-5}$ relative abundances. This yielded a total of 343 species and 417 samples (for 30 subjects, across 16 timepoints). The unique dietary and mid-follow-up antibiotics intervention design of this cohort presented an ideal application opportunity for longitudinal dimension reduction methods, where traditional cross-sectional ordination has limited power to detect the intervention effects (Tanes et al. (2021)).

We represent the structure of such longitudinal microbial abundance datasets with order-3 tensors (Figure 1). Specifically, let the tensor $[X_{ijk}] \in \mathbb{Z}_{+}^{d_M \times d_S \times d_T}$ represent microbiome sequencing counts in a longitudinal study. That is, $X_{ijk}$ is the count of the $i$-th feature, in the $j$-th subject at the $k$-th timepoint. $d_M, d_S, d_T$ indicate the total number of features, subjects, and timepoints respectively. For simplicity, throughout the paper we will use notations such as $X$ and $[X_{ijk}]$ interchangeably to indicate the order-3 tensor with elements $X_{ijk}$. We will also use the term “sample $j,k$” to refer to the sample of subject $j$ measured at timepoint $k$.

The microbial sequencing counts $(X_{1jk}, \ldots, X_{d_M jk})$ target relative abundance of microbial features in sample $j,k$. Formally, let $(p_{1jk}, \ldots, p_{d_M jk})$ denote the relative abundance of microbes in sample $j,k$. Conceptually, these are normalized from actual abundances (e.g. cell counts or concentrations, Vanderputte et al. (2017)); the vector $(p_{1jk}, \ldots, p_{d_M jk})$ thus measures the “relative” abundance of different features and is constrained to sum up to one. The sequencing counts $(X_{1jk}, \ldots, X_{d_M jk})$ is a noisy measurement of the relative abundances. That is, let $N_{jk}$ indicate the total read count (i.e. sequencing depth) for sample $j,k$: $N_{jk} = \sum_{i=1}^{d_M} X_{ijk}$, then (with unbiased bioinformatics protocols) $EX_{ijk} = N_{jk} p_{ijk}$. In common microbiome analysis applications, $p_{ijk}$ are often the primary parameters of interests; for example, our goal of dimension reduction involves detecting low rank structures in $[p_{ijk}]$.

We reiterate the longitudinal microbiome data issues discussed in Section 1 using these notations:

- Sparsity. $X_{ijk}$’s are often zero. This is due to underlying $p_{ijk}$ being either zero or close to zero. In the latter case the sequencing depth $N_{jk}$ through modern bioinformatics is not adequate to detect the given features and will report zero count measurements.

- Compositionality. The relative abundance vector $(p_{1jk}, \ldots, p_{d_M jk})$ is constrained to sum up to one ($\sum_{i=1}^{d_M} p_{ijk} = 1$).

- Overdispersion. Sequencing counts $X_{ijk}$ are highly variable given their mean $N_{jk} p_{ijk}$. Thus restrictive parametric count models (such as the Poisson or multinomial) often needs to be relaxed to allow for large Var($X_{ijk}$’s).

- Longitudinality. Microbial abundances for the same subject at different timepoints $(p_{1jk}, \ldots, p_{d_M jk})$ and $(p_{1jk'}, \ldots, p_{d_M jk'})$ are often associated with each other.


3.1. The mean structure decomposition model. We adapt the canonical polyadic decomposition (Hitchcock (1927)) for dimension reduction on microbiome longitudinal observations (Figure 1). Specifically, this is realized by adapting the CPD low rank decomposition on the mean component (i.e. relative abundance) of microbial count observations $[X_{ijk}]$, with an additional link function:

$$E(X_{1jk}, \ldots, X_{d_M jk}) = N_{jk} \cdot (p_{1jk}, \ldots, p_{d_M jk}),$$

(1)

$$
(p_{1jk}, \ldots, p_{d_M jk}) = \left( \frac{\exp Y_{1jk}}{\sum_{i=1}^{d_M} \exp Y_{ijk}}, \ldots, \frac{\exp Y_{d_M jk}}{\sum_{i=1}^{d_M} \exp Y_{ijk}} \right),
$$

(2)
Fig 1. Graphical representation of our dimension reduction model. Longitudinal microbial measurements of features/subjects over time can be represented as a mode-3 tensor. Potential low rank structures within this tensor is modeled via factorization into rank-1 factors, each decomposed into feature-, subject-, and time-specific loadings.

\[
[Y_{i,j,k}] = \sum_{r=1}^{R} \lambda_r M_r^{d_M} \otimes S_r^{d_S} \otimes T_r^{d_T}.
\]

In this model, \((p_{1,j,k}, \ldots, p_{d_M,j,k}) \in (0,1)^{d_M}\) is the vector of relative abundances for sample \(j,k\) as defined above. \(\mathbf{1}\) essentially claims that the bioinformatics protocols provide unbiased measurements for true microbial relative abundances, and is commonly assumed in microbiome data modelling (Kuczynski et al. (2012)). The relative abundances are further parameterized by \((Y_{1,jk}, \ldots, Y_{d_M,jk}) \in \mathbb{R}^{d_M}\), the vector of centered log ratio (CLR) transformed compositional abundances. That is, \(Y_{i,j,k} = \log \frac{p_{i,j,k}}{g(p_{i,j,k})}\), where \(g(p_{i,j,k}) = (\prod_{i=1}^{d_M} p_{i,j,k})^{\frac{1}{d_M}}\) is the geometric mean of the elements \((p_{1,j,k}, \ldots, p_{d_M,j,k})\). The reverse mapping \(Y_{i,j,k} \rightarrow p_{i,j,k}\) induces \(\mathbf{2}\). The CLR has been previously applied to microbiome data for ease of operations and interpretations, as it transforms constrained compositional abundances into unconstrained Euclidean space (Tsilimigras and Fodor (2016)). Equation (3) is the CPD-inspired dimension reduction model. Here \(\lambda_r \in \mathbb{R}_+\) are singular values and \(M_r, S_r,\) and \(T_r\) are norm-1 “loadings” corresponding to microbial features, subjects, and timepoints. \(||M_r||_2 = ||S_r||_2 = ||T_r||_2 = 1\) per common definition of loading vectors. (3) specifies that the biological signals in the \(d_M \times d_S \times d_T\)-dimensional count data \(\mathcal{X}\) can be decomposed into a total of \(R\) rank-1 factors, each prescribed by the corresponding singular value and singular vectors. The CLR transformation applied above is important for the appropriateness of this model, as \(Y_{i,j,k}\) and the specified decomposition is not constrained into the \(d_M\)-dimensional simplex (unlike \(p_{i,j,k}\)).

This model performs dimension reduction while accounting for data issues in longitudinal microbiome studies. Specifically, dimension reduction is realized by decomposing \(\mathcal{Y}\) (and thus \([p_{i,j,k}]\)) with a total of \(R\) rank-1 tensor factors. Dependency across timepoints of the same subject is characterized via the outer product operation \(S_r \otimes T_r\), just as in regular CDP. Compositionality is naturally satisfied with the parameterization \(\sum_i p_{i,j,k} = 1\). Sparsity in \(X_{i,j,k}\) is accounted for with \(p_{i,j,k}\)’s with values close to zero. We propose in the next subsection a quasi-likelihood based objective function to address overdispersion.

Our model is parameterized by the singular values \(\lambda_r\) and mode-specific loadings \(M_r, S_r,\) and \(T_r\). It is helpful to clarify the interpretations of these parameters. To this end, we note that the relative abundance log-ratio between the features \(i_1\) and \(i_2\) in sample \(j,k\), as specified through \(\mathbf{2}\) and \(\mathbf{3}\) is:

\[
\log \frac{p_{i_1,j,k}}{p_{i_2,j,k}} = \sum_{r=1}^{R} \lambda_r S_{r,j} T_{r,k} (M_{r,i_1} - M_{r,i_2}).
\]
Thus the microbial loading $M_r$ can be viewed as the “basis” for microbial feature abundances in factor $r$; the contrasts between its elements, $M_{ri1} - M_{ri2}$, dictates the difference of feature abundances on the log-ratio scale. $S_r$ and $T_r$ impose subject- and timepoint-specific multiplicative effects on such contrasts; they modify the magnitude and direction of the microbial differences in a per-subject or per-timepoint fashion. Lastly, $\lambda_r$ specifies the overall data variability characterized by factor $r$, since $M_r$, $S_r$, $T_r$ are all norm-1 vectors.

3.2. A semiparametric quasi-likelihood approach for overdispersion. Given the mean assumption (1), it is natural to assume a parametric multinomial distribution on $X_{jk}$, given $p_{jk}$ and $N_{jk}$. However, past investigation in high throughput sequencing has reported notable overdispersion in the data (Love, Huber and Anders (2014)). That is, the observed variability in sequencing counts often far exceeds the expectations prescribed by restrictive parametric models. If not addressed, such overdispersion has can lead to underpowered estimation and incorrect inference (Kvam, Liu and Si (2012)). In our case, a multinomial model implies a restricted mean-variance relationship that is fully determined by the mean parameters $\gamma$: $\text{Var}(X_{ijk}) = N_{jk}p_{ijk}(1 - p_{ijk})$. In practice, we observed severe overdispersion in the real-world FARMM microbial sequencing counts (Figure 2A).

**Fig 2.** Strong overdispersion observed in real-world data is addressed by quasi-likelihood modeling. We fitted decomposition multinomial or quasi-likelihood models to the FARMM dataset (Tanex et al. (2021), details in Section 5). We estimated per-sample “overdispersion factor” in the data, which is defined as the ratio between observed read count variance and model-based, expected variance. This factor quantifies the goodness-of-fit of a given model in terms of read count variances. A) For the multinomial model, the overdispersion factors are heavily inflated above one, and strongly correlates with per-sample sequencing depth, which is commonly observed for microbiome sequencing data (McMurdie and Holmes (2014)). B) Quasi-likelihood modelling accounts for overdispersion, where the model-based overdispersion factors are centered around one. See Supplemental Materials for further explanation.

We thus adopt a quasi-likelihood modification on the model objective function. Specifically, we note the following semiparametric model satisfies the mean assumption (1):

$$X_{jk} \sim \text{Multinomial}(N_{jk}, \tilde{p}_{jk}),$$

$\tilde{p}_{jk}$ is random; $E\tilde{p}_{jk} = p_{jk}$. 

$$\log_{10} \text{Overdispersion Factor}$$

$$\log_{10} \text{Read Depth}$$

$$\log_{10} \text{Overdispersion Factor}$$

$$\log_{10} \text{Read Depth}$$
Additionally, because \( \text{Var}(\tilde{p}_{ijk}) = E(\tilde{p}^2_{ijk}) - (E\tilde{p}_{ijk})^2 \leq E(\tilde{p}_{ijk}) - (E\tilde{p}_{ijk})^2 = p_{ijk}(1-p_{ijk}) \),

there exists \( \phi \in [0, 1] \) such that \( \text{Var}(\tilde{p}_{ijk}) = \phi p_{ijk}(1-p_{ijk}) \). This induces that \( \text{Var}(X_{ijk}) = N_{jk}p_{ijk}(1-p_{ijk}) \times [1 + (N_{jk} - 1)\phi] \). Comparing this against the multinomial variance of \( X_{ijk} \), we note that the semiparametric model allows for a variance inflation factor of \( [1 + (N_{jk} - 1)\phi] \).

To incorporate the improved variance specification for an objective function, we first examine the baseline multinomial likelihood:

\[
I^{(\text{Multinomial})} (\lambda, M, S, T | \mathcal{X}) = \sum_{j,k} \left\{ \sum_i X_{ijk} Y_{ijk} - N_{jk} \log(\sum_i \exp Y_{ijk}) \right\}.
\]

The term for sample \( j, k \) satisfies the exponential-dispersion specification

\[
\frac{Y \theta - b(\theta)}{a(\phi)}
\]

and follows its prescribed mean-variance relationship (\( EX = b'(\theta), \text{Var}(X) = b''a(\phi) \)). To allow for additional variability, we simply replace the current restrictive \( a(\phi) = 1 \) with the more flexible specification derived above: \( a_{jk}(\phi) = [1 + (N_{jk} - 1)\phi] \). This determines that \( \text{Var}(X_{ijk}) = N_{jk}p_{ijk}(1-p_{ijk}) \times [1 + (N_{jk} - 1)\phi] \). The new “quasi-likelihood” objective function is

\[
I^{(\text{Quasi})} (\lambda, M, S, T, \phi | \mathcal{X}) = \sum_{j,k} \left\{ \sum_i X_{ijk} Y_{ijk} - N_{jk} \log(\sum_i \exp Y_{ijk}) \right\}.
\]

The term “quasi” refers to the fact that \( I^{(\text{Quasi})} \) is not a valid likelihood based off a parametric distribution; rather, it is semiparametrically adjusted from the multinomial likelihood, to include the additional variance inflation factor \( a(\phi) = [1 + (N_{jk} - 1)\phi] \). \( \phi \) is an additional parameter beyond the mean parameters of interests (\( \lambda, M, S, T \)) to control for overdispersion. Compared to the fully parametric multinomial likelihood (4), the quasi-likelihood (5) provides significantly improved fit to the variances in real-world microbial observations in FARMM (Figure 2B). We also show in our simulation studies (Section 4) that this leads to estimation efficiencies for the mean parameters, which are of primary interests.

We term our method “microTensor” (tensor decomposition for microbiome data), and evaluate the performance of both the baseline, unweighted multinomial version, and the quasi-likelihood version accounting for overdispersion in our simulation studies.

### 3.3. Parameter constraints.

Before detailing a fitting procedure, we first discuss our model’s parameter constraints, required for either identifiability or interpretability considerations:

- **C1** \( \lambda_r > 0; ||M_r||_2 = ||S_r||_2 = ||T_r||_2 = 1 \).
- **C2** \( \sum_{i=1}^{d_r} M_{ri} = 0 \) for all \( r \).
- **C3** \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_R \).
- **C4** \( \max_j S_{rj} = \max_j |S_{rj}|, \max_k T_{rk} = \max_k |T_{rk}| \).
- **C5** The columns of \( M = (M_1, \ldots, M_r) \) are orthogonal.
- **C6** \( T_{rk} \geq 0 \) for all \( r \) and \( k \).

(C1) is required by our model definition. (C2) is a necessary compositional identifiability constraint. Note that for any constant \( c \), the microbial loading vectors \( (M_{r1}, \ldots, M_{rd_r}) \) and \( (M_{r1} + c, \ldots, M_{rd_r} + c) \) yield the same abundances \( [p_{ijk}] \) according to (2), and thus induce the same likelihood. In other words, \( M_r \) can only be identified up to a constant. On the other
hand, we note that for all sample $j, k$, $\sum_{i=1}^{d_M} Y_{ijk} = \sum_{i=1}^{d_M} \log \frac{p_{ijk}}{g(p_{ijk})} = 0$. To enforce this, we simply require that $\sum_{i=1}^{d_M} M_{ri} = 0$.

(C3) and (C4) are necessary permutation identifiability constraints. (C4) effectively requires that the maximum absolute value elements of each $S_r$ and $T_r$ are non-negative. These conditions ensure that the estimated parameters are uniquely identifiable (otherwise e.g. $(S_r, T_r)$ and $(-S_r, -T_r)$ yield the same likelihood).

(C5) and (C6) are optional interpretability constraints: (C5) enables the decomposition of variability in $\mathcal{Y}$ by its factors; given that $M$ has orthonormal columns, it can be easily shown that $||\mathcal{Y}||_F^2 = \sum_{r=1}^{R} \lambda_r^2$ (here $|| \cdot ||_F$ is the tensor Frobenius norm, i.e., $||\mathcal{Y}||_F^2 = \sum_{i=1}^{d_M} \sum_{j=1}^{d_S} \sum_{k=1}^{d_T} Y_{ijk}^2$). By enforcing (C5) we can provide interpretations for the fitted results that are otherwise inappropriate, including designating percent variability explained for each factor $r$ as in principle component analysis, or visualizing factor loadings against each other on a two-dimensional Euclidean plain as in regular microbiome ordination (McMurdie and Holmes (2013)).

(C6) requires that all elements of the timepoint loadings $T_r$ are non-negative. As we show in our empirical real-world analysis examples, this can lead to the shrinkage of $T_{rk}$ to zeros for $ks$ either before or after a certain critical timepoint during the study period. This can correspond to e.g. the administration of interventions or the development of clinical conditions. Thus, $T_{rk}$’s switching from zeros to positive values (or the opposite) describes the “activation” or “wash-out” of microbial contrasts for factor $r$, due to the effect of treatment or clinical conditions at certain landmark timepoints.

We note that while (C5) and (C6) provide interpretability conveniences, they are introduced solely for such purposes and should not be confused with the necessary identifiability conditions (C1-C4). In fact, we reference the following uniqueness results from De Lathauwer (2006):

**Lemma 3.1.** (De Lathauwer, 2006) Suppose $d_M > d_S$ and $d_M > d_T$ (as is usually the case for microbiome longitudinal studies). Also assume without loss of generality that $d_S > d_T$. If $R \leq d_T$ and $R(R-1) \leq d_M(d_M-1)d_S(d_S-1)/2$, then the rank-$R$ semi-nonnegative tensor decomposition satisfying (C1-C4)

$$\mathcal{Y}^{d_M \times d_S \times d_T} = \sum_{r=1}^{R} \lambda_r M_r^{d_M} \otimes S_r^{d_S} \otimes T_r^{d_T}$$

is unique for almost all such tensors except on a set of Lebesgue measure zero.

DeLathauwer’s original results were derived for unconstrained $\mathcal{Y}$ but can be easily applied to the special case where $\sum_{i=1}^{d_M} Y_{ijk} = 0$ (i.e. our model). This uniqueness conclusion first establishes that our model can be uniquely identified, given the necessary constraints (C1-C4), and second shows that not every $\mathcal{Y}$ with a low rank structure can be decomposed while also satisfying the interpretability constraints (C5) and/or (C6). However, in practice it is often desirable to additionally impose such constraints for more clinically/biologically meaningful fitting results. As discussed in the following Model fitting section, our fitting algorithm is heuristic based and cannot guarantee convergence to the true values (as is common for existing dimension reduction methods for tensor data), with or without the interpretability constraints.

### 3.4. Model fitting

Simultaneous optimization of the quasi-likelihood objective function (5) can be difficult. However, we note that the score function corresponding to baseline multinomial likelihood (4) is still unbiased given the model’s mean assumptions eqs. (1) to (3).
Thus, estimations for the mean parameters $\lambda, M, S, T$ based on optimizing (4) is consistent; they will only be less efficient since the score for sample $j, k$ is non-optimally weighted. This inspires us to take a two-step procedure for optimizing (5): first, we fit the “working” objective function (4) to obtain unbiased estimates for the mean parameters $\lambda, M, S, T$. Second, the mean parameter estimators used to provide a “one-step” estimation $\hat{\phi}$ for the variance parameter $\phi$. Lastly, we plug $\hat{\phi}$ into (5) to obtain final parameter estimates for the mean parameters. In details,

1. Solve for $\hat{\lambda}_{(\text{Multinomial})}, \hat{M}_{(\text{Multinomial})}, \hat{S}_{(\text{Multinomial})}, \hat{T}_{(\text{Multinomial})}$ by optimizing (4).
2. Use $\hat{\lambda}_{(\text{Multinomial})}, \hat{M}_{(\text{Multinomial})}, \hat{S}_{(\text{Multinomial})}, \hat{T}_{(\text{Multinomial})}$ to calculate residuals and estimate $\hat{\phi}$. Specifically:
   - Under our model assumptions, $E(X_{ijk} - N_{jk}p_{ijk})^2 = N_{jk}p_{ijk}(1 - p_{ijk})(1 + (N_{jk} - 1)\phi]$.
   - Summing over microbial features for each subject/timepoint and rearranging terms, this gives $\phi = E(\sum(N_{jk}p_{ijk}(1 - p_{ijk})) - 1)/(N_{jk} - 1)$.
   - $\phi$ is thus estimated with $\hat{\phi} = \frac{1}{d_{S}d_{T}} \sum_{j,k}
   \left(\frac{\sum(X_{ijk} - N_{jk}p_{ijk})^2 - 1}{N_{jk} - 1}\right)$, where $\hat{p}_{ijk}$ is obtained by plugging $\hat{\lambda}_{(\text{Multinomial})}, \hat{M}_{(\text{Multinomial})}, \hat{S}_{(\text{Multinomial})}, \hat{T}_{(\text{Multinomial})}$ into (3) and (2).
3. Use $\hat{\phi}$ to reweight observations and optimize (5) to obtain improved estimations $\hat{\lambda}_{(\text{Quasi})}, \hat{M}_{(\text{Quasi})}, \hat{S}_{(\text{Quasi})}, \hat{T}_{(\text{Quasi})}$.

Because $\hat{\phi}$ is already a consistent estimator of $\phi$, additional steps iterating over $(\hat{\lambda}_{(\text{Quasi})}, \hat{M}_{(\text{Quasi})}, \hat{S}_{(\text{Quasi})}, \hat{T}_{(\text{Quasi})})$ and $\hat{\phi}$ only marginally improves the efficiency in estimating $\lambda, M, S, T$, the primary parameters of interest. This is discussed in more details in the Supplemental Materials.

Given known score weights, we still need to optimize the objective functions with respect to the mean parameters. To this end, we follow the proposed practice in Wang, Fischer and Song (2019), and adopt a greedy procedure to solve for each factor $r$ successively. That is, we solve for $(\hat{\lambda}_{r}, \hat{M}_{r}, \hat{S}_{r}, \hat{T}_{r})$ via:

(6)

\[
\text{minimize}_{\lambda_{r}, M_{r}, S_{r}, T_{r}} - l^{(r)}(\lambda_{r}, M_{r}, S_{r}, T_{r} | X, \hat{\mathcal{Y}}^{(r-1)}) = \\
- \sum_{j,k} \left\{ \sum_{i} X_{ijk}(\hat{Y}_{ijk}^{(r-1)} + \lambda_{r} M_{r} \otimes S_{r} \otimes T_{r}) - N_{jk} \log(\sum_{i} \exp(\hat{Y}_{ijk}^{(r-1)}) + \lambda_{r} M_{r} \otimes S_{r} \otimes T_{r}) \right\}
\]

subject to constraints (C1-C4) and optionally (C5-C6), where $\hat{\mathcal{Y}}^{(r-1)}$ is the approximated $\mathcal{Y}$ from model fitting of the first $r - 1$ factors; $\hat{\mathcal{Y}}^{(0)} = 0^{d_{S} \times d_{T}}$, and $\hat{\mathcal{Y}}^{(r)} = \sum_{l=1}^{r} \lambda_{r} M_{r} \otimes S_{r} \otimes T_{r}$ for $r \geq 1$. That is, during each factor’s fitting step, we aim to maximize the model likelihood by adding the additional $r$-th factor to the existing $(r - 1)$-factors estimation $\hat{\mathcal{Y}}^{(r-1)}$. $a_{ijk}(\hat{\phi})$ is the weighting function (either constant 1 for (4) or 1 + $(N_{jk} - 1)\phi$ for (5)).

The minimization problem (6) is non-convex in $\lambda_{r}, M_{r}, S_{r}, T_{r}$, and has non-convex constraints (C1) ($||M_{r}|| = ||S_{r}|| = ||T_{r}|| = 1$). However, we note that a) the objective is component-wise convex for each of $\lambda_{r}, M_{r}, S_{r}, T_{r}$, allowing us to adopt a coordinate-based approach, and b) the non-convex constraint (C1) can be satisfied by fitting an unconstrained optimization first, and rescaling the fitted parameters $(\hat{\lambda}_{r}, \hat{M}_{r}, \hat{S}_{r}, \hat{T}_{r})$ to
\( (\hat{\lambda}_r, ||\hat{M}_r||_2, ||\hat{S}_r||_2, ||\hat{T}_r||_2, \frac{||\hat{M}_r||_2}{||\hat{S}_r||_2}, \frac{||\hat{S}_r||_2}{||\hat{T}_r||_2}) \). This satisfies the constraint without changing the value of the optimized objective function. We thus propose the following coordinate-wise projected gradient descent algorithm to solve (6):

**Algorithm 1:** Coordinate gradient descent to solve for factor \( r \)

**Input:** Microbiome count tensor \( \mathcal{X} \), estimated \( \mathcal{Y} \) from the first \((r - 1)\) factors, \( \hat{\mathcal{Y}}^{(r-1)} \)

**Output:** Parameter estimates \( \hat{\lambda}_r, \hat{M}_r, \hat{S}_r, \hat{T}_r \)

Initialize \( \hat{\lambda}_r^{(0)}, \hat{M}_r^{(0)}, \hat{S}_r^{(0)}, \hat{T}_r^{(0)} \) with classical CPD;

**while not converged do**

During step \( l \):

- Solve for \( \hat{M}_r^{(l)} \) by optimizing \( -l^{(r)}(\lambda_r^{(l-1)}, M_r^{(l-1)}, S_r^{(l-1)}, T_r^{(l-1)}|X, \hat{Y}^{(r-1)}) \) with projected gradient descent, subject to constraint (C2) and optionally constraint (C5).
- Solve for \( \hat{S}_r^{(l)} \) by optimizing \( -l^{(r)}(\lambda_r^{(l-1)}, M_r^{(l)}, S_r^{(l)}, T_r^{(l)}|X, \hat{Y}^{(r-1)}) \) with gradient descent.
- Solve for \( \hat{T}_r^{(l)} \) by optimizing \( -l^{(r)}(\lambda_r, M_r^{(l-1)}, S_r^{(l)}, T_r^{(l)}|X, \hat{Y}^{(r-1)}) \) with (projected) gradient descent, optionally subject to constraint (C6).
- Solve for \( \hat{\lambda}_r^{(l)} \) by optimizing \( -l^{(r)}(\lambda_r, M_r^{(l)}, S_r^{(l)}, T_r^{(l)}|X, \hat{Y}^{(r-1)}) \) with gradient descent.

**end**

Flip the signs of converged \( \hat{M}_r, \hat{S}_r, \hat{T}_r \) if necessary to satisfy constraint (C4);

Lastly, replace \( \hat{\lambda}_r, \hat{M}_r, \hat{S}_r, \hat{T}_r \) with \( (\hat{\lambda}_r, ||\hat{M}_r||_2, ||\hat{S}_r||_2, ||\hat{T}_r||_2, \frac{||\hat{M}_r||_2}{||\hat{S}_r||_2}, \frac{||\hat{S}_r||_2}{||\hat{T}_r||_2}) \) to satisfy constraint (C1).

The algorithm is guaranteed to converge to a solution, as the objective function has a lower bound, and each of the (projected) gradient descent steps decreases its value. The last two operations modify the fitted parameter values without changing the optimized objective function value.

We make three additional comments regarding our model fitting. First, our model easily handles missing timepoint measurements that are common in microbiome longitudinal studies (Pasolli et al. (2017)). This is because the form of the objective functions (4) and (6) take outer summations over subject and timepoint indices \( j \) and \( k \). This can be modified to include only measured timepoints to accommodate missing data. Second, we note again neither the successive optimization for each factor \( r \), nor the inner optimization (6) via coordinate gradient descent guarantees convergence to the true parameter values. We demonstrate with simulation studies and real-world applications that, empirically, our fitting procedure lead to clinically meaningful results that are close to the truth (for simulation studies) or observed data (for real-world studies). Last, in practice, the number of true factors \( R \) is unknown and must be estimated from data. This is a common issue in factorization methods that is often addressed by heuristics, such as examining the residual sum of squares at different numbers of total factors (Wang, Fischer and Song (2019)). Given our model specification, we recommend fitting with different \( R \)'s and comparing the model negative log-likelihood. An “elbow” point in the negative log-likelihood curve provides evidence for selecting the appropriate \( R \) (Supplemental Figure 1).

4. Simulation. We conducted three simulation studies to evaluate the performance of our method, microTensor, when compared against existing ordination dimension reduction methods for correlated microbiome data. The first two studies evaluated longitudinal microbiome decomposition models, namely, microTensor and CTF (Martino et al. (2021)), in their accuracy to recover the original relative abundance tensor \([p_{ijk}]\). True tensor parameters \([p_{ijk}]\)
were generated either according to our model (4.1, correctly specified model) or that of CTF (4.2, misspecified model). A variety of parameter settings, including noise level, read depth, data missingness, and number of independent subjects were considered. In both scenarios microTensor outperformed CTF, for the vast majority of parameter combinations.

In the third scenario, we adopted an existing microbiome simulation software (Ma et al. (2021)) and performed simulations by perturbing an empirically fitted real-world dataset to incorporate within-subject correlations and case-versus-control disease effects, without imposing low rank structures during data generation. The goal is to determine different methods’ statistical power in detecting disease effects in their top discovered subject loadings, importantly in more realistic settings where low rank structures cannot be assumed a priori. We compared microbiome-specific dimension reduction methods, microTensor and CTF, and additionally the naive principal component analysis, and observe that microTensor was the most powerful in detecting case-versus-control differences through its identified loadings.

4.1. Correctly specified model. For this simulation study, a two-factor microbial longitudinal data tensor was simulated, with 198 microbes, 100 subjects, and 10 time points. Specifically, we have:

\[ Y = \lambda_1 M_1 \otimes S_1 \otimes T_1 + \lambda_2 M_2 \otimes S_2 \otimes T_2 \]

where the first factor corresponded to common microbial composition community structure, with \( \lambda_1 = 1000 \) and

\[ M_1 \propto (\begin{array}{ccc}
\underline{1,1,\cdots,1} & \underline{0,0,\cdots,0} & \underline{-1,-1,\cdots,-1}
\end{array})
\]

66 high abundance features 66 medium abundance features 66 low abundance features

\[ S_1 \propto (1,1,\cdots,1), \quad T_1 \propto (1,1,\cdots,1)
\]

100 subjects 10 timepoints

This factor corresponded to a microbial compositional structure that was common across subjects and timepoints, and differentiated between groups of high, medium, and low abundance features. Note that \( M_1, S_1, T_1 \) in practice were normalized to norm-1; here we present them in a scaled form for ease of illustration.

The second factor corresponded to the development of microbial perturbations between two groups of subjects across time, with \( \lambda_2 = 500 \) and

\[ M_2 \propto (\begin{array}{ccc}
\underline{1,0,-1,\cdots,1,0,-1} & \underline{1,0,-1,\cdots,1,0,-1} & \underline{1,0,-1,\cdots,1,0,-1}
\end{array})
\]

contrasts in high abundance features contrasts in medium abundance features contrasts in low abundance features

\[ S_2 \propto (0,0,\cdots,0,1,1,\cdots,1), \quad T_2 \propto (\begin{array}{ccc}
\underline{0,0,\cdots,0} & \underline{1,1,\cdots,1}
\end{array})
\]

50 “controls” 50 “cases” 5 pre-“disease” timepoints 5 post-“disease” timepoints

This corresponded to the development of a potential “disease” condition between timepoints five and six, in the 50 “case” subjects, that lead to further microbial abundance differentiation within the low, medium, and high abundance groups of microbes.

Given \( Y \), we obtained the per subject/timepoint relative abundances \( (p_{1jk}, \ldots, p_{dM,jk}) \) based on (2). We note that such parameter specifications yielded a range of \( p_{ijk} \) that can be expected for real-world data, i.e., \( 10^{-6} \sim 10^{-1} \) (Supplemental Figure 2). Given \( [p_{ijk}] \), per-sample simulated counts were generated with a overdispersed Dirichlet-multinomial distribution,

\[ (X_{1jk}, \ldots, X_{dM,jk}) \sim \text{Dirchlet-multinomial}(N_{jk}, \alpha_0 p_{1jk}, \ldots, \alpha_0 p_{dM,jk}). \]
Equivalently, \( \tilde{p}_{ijk} \) were generated with a Dirichlet distribution with parameters \( p_{ijk} \) and \( \alpha_0 \), and \( X_{ijk} \) were generated as multinomials with read depth \( N_{jk} \) and relative abundances \( \tilde{p}_{ijk} \). We used the Dirichlet-multinomial distribution to allow for additional biological/technical noise in the observed counts. With this process we have that \( E X_{ijk} = N_{jk} p_{ijk} \), i.e., the measured relative abundance for each feature was unbiased for the underlying truth.

However, the variability for \( X_{ijk} \) was higher than that expected of a standard multinomial distribution: \( \text{Var}(X_{ijk}) = N_{jk} p_{ijk}(1 - p_{ijk}) \frac{N_{jk} + \alpha_0}{1 + \alpha_0} \). This accounted for the possibility that the “measured” relative abundance can fluctuate given the underlying proportions, due to biological or technological noises. We note that this variability term agreed with that specified through our quasi-likelihood model: \( a_{jk}(\phi) = 1 + (N_{jk} - 1) \frac{1}{1 + \alpha_0} \) where \( \phi = \frac{1}{1 + \alpha_0} \). The \( \alpha_0 \) and \( \phi \) parameters both characterized overdispersion, or the strength of noises: smaller \( \alpha_0 \)’s corresponded to higher variances, and thus larger noise effects and more overdispersed \( X_{ijk} \). The Dirichlet-multinomial converges to the multinomial distribution as \( \alpha_0 \to \infty \).

The list of parameters varied in this simulation study were:

- The noise parameter \( \alpha_0 \), ranging from \( 10^{2.5} \) to \( 10^4 \) on an equally-spaced log scale.
- Median of per-sample read depth \( N_{jk} \) was set at either 10,000 or 100,000, mimicking modern sequencing studies. E.g. for median depth of 10,000 reads, \( \log_{10} N_{jk} \sim \text{Uniform}(3, 5) \).
- Percentage of missing observations. For each subject, either all observations were observed, or four timepoints (40%) were simulated to be missing at random.

For each parameter combination, 1,000 random Monte Carlo simulation datasets \([X_{ijk}]\) were generated. In addition to the realistic range of true relative abundances \( p_{ijk} \), the above parameter choices also yielded realistic data sparsity and a reasonable variability for the noise level in \( \tilde{p}_{ijk} \), when compared to the true parameters \( p_{ijk} \) (Supplemental Figure 2).

We fitted CTF and microTensor (both the quasi-likelihood and the unweighted multinomial version) to the simulated datasets. For method evaluation metrics, we note that both microTensor and CTF provide point estimates for the true relative abundance parameters \( \tilde{p}_{ijk} \). This can be obtained by plugging the identified factors \( \lambda_r, M_r, S_r, T_r \) into either method’s tensor decomposition model, and then taking the expit transformation (2). We thus adopted (log-transformed) relative \( L_1 \) difference between \( \tilde{p}_{ijk} \) and true \( p_{ijk} \), to evaluate the methods’ accuracy in capturing the true underlying relative abundance tensor \([p_{ijk}]\):

\[
\sum_{ijk} \log_{10} \sum_i \frac{|p_{ijk} - \tilde{p}_{ijk}|}{p_{ijk}} d_S d_T
\]

The performance comparison results are summarized in Figure 3. Our model consistently outperformed CTF, and more so under large variability settings (small \( \alpha_0 \)). There are two reasons: first, the true simulation model indeed agreed with our approach, and thus less bias was expected of our method (the CTF model is detailed in the following subsection). Second, the generalized centered log-ratio transformation in CTF specifies its per-sample normalization factor as the geometric mean of all non-zero feature abundances. In practice, the non-zero feature set varies from sample to sample, especially under higher sparsity, inducing loss of efficiency compared to microTensor. The problem is further exemplified when we compared the relative \( L_1 \) difference on a per-feature level instead of aggregating across all features, where for features with the highest level of sparsity microTensor outperformed CTF the most (Supplemental Figure 3). This was further validated in the following two scenarios, where the simulation model did not agree with our approach. We also observe that the quasi-likelihood version of our method (microTensor) outperformed the multinomial version (microTensor (uw)). This is due to the improved per-observation weighting in the quasi-likelihood that better accounted for overdispersion in the data.
Fig 3. microTensor had improved performance in dimension reduction for microbiome count observations in simulation. microTensor and CTF were compared in terms of the relative $L_1$ difference between the estimated and true relative abundance tensor. Error bars indicate one standard error of the estimated $L_1$ difference at each parameter combination, across the 1,000 Monte Carlo repetitions. We additionally varied the data overdispersion, sequencing depth, and missing samples per-subject. Smaller $\alpha_0$ indicates higher overdispersion. Both the unweighted (microTensor (uw)) and quasi-likelihood (microTensor) versions of our models performed better than CTF, while the quasi-likelihood model further improved accuracy especially for heavily overdispersed data.

Lastly, we compared the computational time of microTensor versus CTF. CTF by design is faster as its coordinate updates have analytical solutions and do not require numerical optimization. For better computation speed, we implemented the core gradient descent in microTensor with C++. As presented in Supplemental Figure 4, microTensor required more computation time than CTF but is not prohibitively so, with each dataset fitting finished under one minute.

4.2. Misspecified model. In this setting, $Y$ had exactly the same structure as specified in the correctly specified model, with the same set of factor parameters ($\lambda_r, M_r, S_r, T_r, r \in \{1, 2\}$), and $[p_{ijk}]$ was generated again with the expit transformation (2). However, instead of having $E\tilde{p}_{ijk} = \tilde{p}_{ijk}$ as with the Dirichlet distribution, we first generated random $\tilde{Y}_{ijk}$ with $E\tilde{Y}_{ijk} = EY_{ijk}$,

$$\tilde{Y}_{ijk} = Y_{ijk} + \epsilon_{ijk}$$

where $\epsilon_{ijk} \overset{i.i.d}{\sim} \text{Normal}(0, \sigma^2)$. We then generated $\tilde{p}_{ijk}$ with the expit transformation given $\tilde{Y}$: 

$$\tilde{p}_{ijk} = \frac{\exp \tilde{Y}_{ijk}}{\sum_l \exp \tilde{Y}_{ijk}}$$. 

$X_{ijk}$ was simulated again as multinomials given read depth $N_{jk}$ and
relative abundance $\tilde{p}_{ijk}$. We note that the approach of CTF is modified after the CDP. For both, the squared loss function corresponds directly to an underlying Gaussian assumption on the noise in $\tilde{y}$. Thus the specification here agreed with the true model of CTF. On the other hand, due to non-collapsibility the expit transformation, microTensor had misspecified model ($E\tilde{p}_{ijk} \neq p_{ijk}$).

For simulation parameters, we evaluated different median read depths and data missingness as in 4.1. We additionally varied the standard error parameter $\sigma$ for Gaussian noise $\varepsilon_{ijk}$, which, as $\alpha_0$ in 4.1, controlled overdispersion (larger $\sigma$ corresponded to more severe overdispersion). We ranged $\sigma$ equally spaced from 0.25 to 2.25. 1,000 Monte Carlo simulations were again performed for each parameter combination. With these parameter specifications, the true relative abundances $p_{ijk}$, overdispersed relative abundances $\tilde{p}_{ijk}$, and simulated microbial read counts had distribution characteristics agreeing with real-world data, and also in concordance with the first simulation scenario (Supplemental Figure 5). We again used the relative $L_1$ difference as the evaluation metric, and compared the true relative abundance tensor parameters $p_{ijk}$ against estimated $\tilde{p}_{ijk}$, calculated from the discovered low rank factors.
Despite working under a misspecified scenario, microTensor still outperformed CTF in most parameter settings (Figure 4). Same as the correctly specified setting, this is due to the high variability introduced by the CTF’s generalized centered log ratio transformation. Only when data sparsity is low, i.e., under higher read depths and smaller data overdispersion, did we observe a favorable performance of CTF compared to microTensor (Supplemental Figures 5). In such settings, the bias in microTensor estimated \( \hat{p}_{ijk} \)’s (due to model misspecification) outweighed the inefficiency of CTF. Our evaluated parameter grid was selected such that the simulated data had characteristics mimicking that of real-world data (Supplemental Figure 5). microTensor still outperformed CTF for features with the highest levels of sparsity (Supplemental Figure 6). We also note that, despite working under a misspecified setting, empirically the quasi-likelihood re-weighting still improved the performance of microTensor. Together, the correctly and misspecified model evaluations provided evidence for the improved performance of microTensor in practical applications.

4.3. Real-data based simulation. Different from the first two simulation scenarios, this scenario compared method performance with data simulated directly from real-world samples. Thus, importantly, low-rank structures were not imposed a priori. We adopted SparseDOSSA (Ma et al. (2021)), a published and validated microbiome simulation method. Briefly, SparseDOSSA fits a penalized multivariate Gaussian copula model with zero-inflated marginal distributions on joint microbial abundance profiles. It then uses the estimated model parameters to simulate synthetic microbial count observations that closely resemble the fitted real-world dataset. It also perturbs the simulated abundances to introduce artificial associations between certain microbial features and metadata variables, inducing for example “disease” or “longitudinal” effects. Given that SparseDOSSA has been validated to generate realistic microbial abundance profiles (Ma et al. (2021)), we adopted this neutral simulation model that did not impose any low rank structures on the simulated data.

Using SparseDOSSA, we evaluated dimension reduction methods’ ability to detect “disease” effects present in longitudinal microbiome studies, i.e., if the top factor loadings discovered by each method can differentiate between case and control subjects. We generated SparseDOSSA simulation datasets that approximated a previously published healthy human gut study (Lloyd-Price et al. (2017)). Using the method’s functionality to introduce artificial microbiome-metadata associations, we perturbed the simulated microbiomes to induce effects (at varying strengths) of a binary “disease” variable (50% “cases” and 50% “controls”). Ten microbial species each were perturbed to be either positively or negatively correlated with the cases, selected from the top differentially abundant features related to Crohn’s disease reported in Lloyd-Price et al. (2019), a longitudinal study of inflammatory bowel diseases. To include longitudinal within-subject correlations, for each simulation dataset we randomly selected twenty microbial species, and each was correlated with a standard normal, subject-specific “random effect” variable. This introduced “longitudinal” effects for the twenty random microbial species (same number as those correlated with the disease variable).

The list of parameters (fixed or varying) used in this study included:

- Number of features \( d_M = 200 \) (most abundant ones in Ma et al. (2021)).
- Number of subjects \( d_S \) at either 100 (50 “cases”, 50 “controls”) or 30 (15 “cases”, 15 “controls”).
- Number of timepoints \( d_T = 10 \).
- Twenty features, selected from Lloyd-Price et al. (2019), were perturbed to introduce effects of the “disease” variable. The effect size of the disease variable varies from 0 to 1.5 equally spaced. This effect size, as prescribed by SparseDOSSA, are log fold changes in non-zero relative feature abundance and log odds ratios in feature prevalence (Ma et al. (2021)).
• Twenty features, selected at random for each simulation dataset, were perturbed to introduce “longitudinal effects”. This is realized by using SparseDOSSA to introduce associations between each feature with a per-subject, standard normal “random effect” metadata variable. The association effect size, with the same interpretable above, is fixed at 1.
• Median read depths at 10,000 and 100,000.

With such simulation specifications, we introduced both biological case-versus-control signals and longitudinal within-subject correlation effects in the simulated datasets. These are thus appropriate for test the competing methods’ performance in terms of identifying the biological signals in the presence of subject-to-subject variability. For each parameter combination, 1,000 Monte Carlo simulation datasets were created. We fitted both microTensor and CTF, as well as the vanilla PCA (on log-transformed relative abundances). As the goal is to detect the simulated disease effects in each method’s discovered top factors, we evaluated if the top two subject loadings ($S_1$ and $S_2$) differentiate between cases and controls. Formally, the subject loadings were tested against the disease variable using simple linear regression. The two $p$-values for significant associations in $S_1$ and $S_2$ were corrected with Bonferroni correction (i.e., multiplied by 2). The PCA does not impose longitudinal structures a priori, and generates instead sample-wise PC scores. We thus evaluated if the top two PC scores detected disease difference, by fitting linear mixed effect models between the PC scores and the disease variable, with subject-specific random intercepts. The two $p$-values for scores 1 and 2 were again corrected with Bonferroni correction.

We used power of detecting disease effects in at least one of the top two loadings (at 0.05 significance level with Bonferroni correction) as the evaluation metric. microTensor consistently outperformed both PCA and CTF in terms of power to differentiate between cases and controls in its top two loadings, across different disease effect sizes, read depths, and number of subjects (Figure 5, Supplemental Figure 7). PCA underperformed compared to CTF, as it did not account for the longitudinal nature of the data. Re-weighting with quasi-likelihood empirically improved the performance of microTensor. We emphasize that a) again, the simulation in this study was performed by a model-neutral method that does not assume a priori low rank structures, and was validated to closely approximate real-world microbiome data, and b) evaluation was set to a very common task in unsupervised dimension reduction, i.e., detecting certain effects in the identified top loadings. These results thus provided strong evidence for the practical utility of microTensor in dimension reduction of longitudinal microbiome studies.

4.4. Simulation results under smaller sample sizes. We also performed similar simulation comparisons when the sample sizes are small and more comparable to our real data analysis in Section 5. We kept all the simulation parameters the same with only $S_1$ and $S_2$ being changed to length of 30, which corresponds to 30 total subjects (15 cases and 15 controls). These new results are included in Supplemental Figures 8, 9, and 10. The conclusions remain the same. We observed that for Simulation 4.1 (correctly specified model) and 4.2 (mis-specified model), microTensor still outperformed CTF. For simulation 4.3, both microTensor and CTF were preferable to vanilla PCA, while microTensor performed the best. Across all scenarios, quasi-likelihood re-weighting improved microTensor’s performance.

5. Real-world data analysis.

5.1. FARMM study. We analyzed the motivating FARMM cohort with three decomposition methods: PCA, CTF, and microTensor. PCA was performed on log transformed relative abundances, whereas CTF and microTensor were performed directly on the longitudinal microbial counts tensor. microTensor was fitted with the optional interpretability constraints
(C5-6). Because both CTF and microTensor requires initialization based on the canonical polyadic decomposition with random starts, we evaluated five random fits for both methods, and adopted the best fit (based on the methods’ objective functions) for each as the final result. We assessed the top two factors of the fitted results (three factors were fitted for CTF and microTensor; the top two factors accounted for > 70% variability for all three methods).

For these real-world results, as the ground truth is unknown, we adopted two criteria to evaluate different methods’ performance. First, as the methods are designed for unsupervised dimension reduction, they should provide a faithful representation of the data with their top factors. To this end, we compared the similarity between observed microbial abundance profiles in the data, versus those reconstructed by the top loadings of each method. This is measured by the per-sample $L_1$ difference:

$$\frac{1}{d_M} \sum_i |p_{ijk}^{\text{Obs}} - p_{ijk}^{\text{Model}}|$$

Where $p_{ijk}^{\text{Obs}}$ is the observed microbial relative abundance: $p_{ijk}^{\text{Obs}} = \frac{X_{ijk}}{\sum_i X_{ijk}}$. $p_{ijk}^{\text{Model}}$ is the reconstructed microbial profiles from PCA, CTF, or microTensor, via their top factors and through the transformations in (3) and (2). Here the metric is on the absolute scale as opposed to the relative scale in simulation studies 4.1 and 4.2, because the observed proportions $p_{ijk}^{\text{Obs}}$ can potentially be zero. These per-sample differences were visualized in Figure 6A, where we observe that our method achieved the smallest $L_1$ difference among the three methods. CTF had the worst performance in this cohort; this is also observed in its top loading microbial features which were not typical representatives of the gut environment (Supplemental Figure 11). We conclude that microTensor provided the best low-dimensional representation of the FARM dataset, through its top factors.

Second, we also show that the top loadings identified by microTensor provided important biological insights, namely, the effects of dietary groups and antibiotics treatment across the
study period (Figure 6). Specifically, the first factor’s feature loading characterized elevated abundances of pathogenic features such as Escherichia coli as well as antibiotics resistant Veillonella species (Figure 6B left panel). These features have been previously associated with antibiotics exposure in diseased, dysbiotic microbiomes (Gowers et al. (2014)). Here we observed the same associations by visualizing the “per-sample” loadings identified by microTensor. Specifically, the loading for sample \( j, k \) was calculated as \( \lambda_j S_{j,k}^T \). These were then summarized separately per dietary group to generate time trajectories (Figure 6B right panel). We note the first factor’s time trajectory across samples was elevated following antibiotics treatment starting at day seven, corresponding to an increase in the aforementioned antibiotics-associated microbes.

The second factor characterized dietary effects before antibiotics intervention. This is most notably present in the top positive loading for Prevotella copri (Figure 6C left panel), previously reported to be associated with non-western, more fiber-enriched diets (Tett et al. (2019)). Correspondingly, the vegan diet group had elevated per-sample loadings compared to the EEN and omnivore groups pre-antibiotics treatment (Figure 6C right panel). This suggests that the fiber-rich vegan diet subjects had increased \( P. \) copri abundances, consistent with previous findings. Such contrasts, however, were strongly reduced by antibiotics, suggesting a universal “wash-out” effect of the treatment. To formally test for the these diet and antibiotics effects in factors 1 and 2, we performed the following linear regression analyses on subject and time loadings:

\[
S_{rj} \sim \beta_{r0} + \beta_{r1} \text{Vegan}_j + \beta_{r2} \text{EEN}_j
\]

\[
T_{rk} \sim \gamma_{r0} + \gamma_{r1} \text{Day}_k I_{0 \leq \text{Day}_k \leq 5} + \gamma_{r2} \text{Day}_k I_{6 \leq \text{Day}_k \leq 8} + \gamma_{r3} \text{Day}_k I_{9 \leq \text{Day}_k \leq 15}
\]
Here \( r \in \{1, 2\} \). The first model is fitted on a per-subject basis, where Vegan\(_j\) and EEN\(_j\) are indicator variables for the corresponding dietary groups, aimed at identifying contrasts against the omnivore diet (\( \beta_{11} \) and \( \beta_{22} \)). The second model is a linear spline fitted on a per-timepoint basis, where \( \gamma_{r1}, \gamma_{r2}, \) and \( \gamma_{r3} \) correspond to the per-day time loading change before, during, and after antibiotics treatment. We are primarily interested in examining the antibiotics treatment effect, \( \gamma_{r2} \). Testing results are as summarized in Table 1, microTensor was able to detect statistically significant effects in both factors for the vegan versus omnivore difference, as well as for the antibiotics treatment effect during days 6, 7, and 8. These results thus agree with Figure 6.

<table>
<thead>
<tr>
<th>Effect</th>
<th>Factor</th>
<th>Model term</th>
<th>Coefficient</th>
<th>Standard error</th>
<th>T statistic</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vegan vs. Omnivore</td>
<td>Factor 1</td>
<td>( \beta_{11} )</td>
<td>-30.3</td>
<td>8.38</td>
<td>-3.62</td>
<td>1.96e-03</td>
</tr>
<tr>
<td></td>
<td>Factor 2</td>
<td>( \beta_{21} )</td>
<td>77.1</td>
<td>28.2</td>
<td>2.73</td>
<td>0.0138</td>
</tr>
<tr>
<td>EEN vs. Omnivore</td>
<td>Factor 1</td>
<td>( \beta_{12} )</td>
<td>-22.1</td>
<td>11.3</td>
<td>-1.96</td>
<td>0.0659</td>
</tr>
<tr>
<td></td>
<td>Factor 2</td>
<td>( \beta_{22} )</td>
<td>4.15</td>
<td>7.88</td>
<td>0.527</td>
<td>0.604</td>
</tr>
<tr>
<td>Daily change (day 6-8)</td>
<td>Factor 1</td>
<td>( \gamma_{21} )</td>
<td>0.0351</td>
<td>9.42e-03</td>
<td>3.72</td>
<td>2.92e-03</td>
</tr>
<tr>
<td></td>
<td>Factor 2</td>
<td>( \gamma_{22} )</td>
<td>-0.0972</td>
<td>7.87e-03</td>
<td>-12.4</td>
<td>3.5e-08</td>
</tr>
</tbody>
</table>

Lastly, we note that PCA or CTF were equivalently or less able to detect such effects. The feature and sample-specific loadings identified PCA also visually corresponded with certain antibiotics and diet-induced microbial shifts (Supplemental Figure 11A). However the vegan and omnivore differences were not successfully detected (Supplemental Table 1). The top features identified through CTF were not typical representative species of the gut environment, as evidenced by the method’s poor fit on the dataset (Figure 6A), thus the corresponding results remain difficult to interpret (Supplemental Figure 11B). Overall, we conclude that microTensor outperformed existing dimension reduction methods in these longitudinal microbial profiles, by a) providing the best representation of the observe data via dimension reduction, and b) detecting strong, statistically significant dietary and antibiotics effects in its identified factors.

5.2. DIABIMMUNE study. The DIABIMMUNE dataset (Yassour et al. (2016)) followed 39 infants (4 c-section, 35 vaginally delivered) from the 2nd month after birth over the first three years of life with monthly fecal sampling. We selected this study to detect the delivery mode effect on infant gut microbiomes with microTensor, and to compare against findings identified by other dimension reduction methods (Martino et al. (2021) studied the same dataset with CTF). For data pre-processing, OTUs were fist aggregated at the genus level. Genera present in no more than ten samples (\( \sim 1\% \)) with at least \( 10^{-5} \) relative abundance were again removed. Both CTF and microTensor were fitted on the microbial count tensor, whereas PCA was fitted on log transformed relative abundances. microTensor was fitted with the optional interpretability constraints (C5-6). Lastly, five random iterations for microTensor and CTF were performed each to select the optimal fit.

We first note that, again, microTensor outperformed both alternatives in reconstructing the microbial profiles in the data through its top factors (Figure 7A), as evidenced by the \( L_1 \) difference between observed and reconstructed relative abundances (\( \hat{p}_{ijk}^{\text{Obs}} \) versus \( \hat{p}_{ijk}^{\text{Model}} \)). In these data, CTF also achieve better performance than PCA, echoing the findings in Martino et al. (2021).

For biological signals, we focused on detecting vaginal versus c-section delivery model difference in the first identified factor, as both CTF and microTensor assigned \( > 90\% \) variabilities for the first identified factor (consistent with Martino et al. (2021)). microTensor
identified positive associations between commensal features with vaginal delivery when compared against c-section infants, echoing previous findings (Shao et al. (2019)). Specifically, microTensor detected high positive loadings for genera such as Bacteroides, Faecalibacterium, and Bifidobacterium in the first factor (Figure 7B), previously reported to be associated with vaginal delivery when compared with c-section (Shao et al. (2019)). Enrichment for these features were elevated in vaginally delivered infants, as visualized through subject loadings (Figure 7B). The difference in microTensor identified subject loadings of between c-section and vaginally delivered infants was statistically significant (two sample t-test p = 0.029).

The results of microTensor and CTF largely agreed (Supplemental Figure 12B). CTF also assigned high feature loadings to features such as Bacteroides and Faecalibacterium. It also visually detected differences between the c-section and vaginal delivery subject loadings, albeit without statistical significance (two sample t-test p = 0.380). The two methods also agreed that strongest variation in the DIABIMMUNE cohort corresponded delivery mode effects (first factor > 90% variability). This was consistent with the original evaluation reported in Martino et al. (2021). PCA, in contrast, assigned much less variance to the first factor, potentially due to the confounding effect of the longitudinal design, and was unable to detect difference in sample loadings (Supplemental Figure 12A, two sample t-test p = 0.425). Lastly, no strong time effects could be identified from microTensor’s first loading (Supplemental Figure 13). Overall, we conclude that for the DIABIMMUNE cohort, microTensor provided the best low-dimension representation of the data that closely approximate the observed microbiomes; CTF and microTensor agreed on the strong delivery mode effect on infant gut microbiomes that vanilla PCA failed to detect.

6. Discussions. In this work, we introduced microTensor, a tensor factorization based method for ordination dimension reduction of microbiome data. Model wise, microTensor accounts for the data-specific characteristics of longitudinal microbiome cohorts, including within subject correlations, sparsity, compositionality, and overdispersion. We provide a quasi-likelihood based objective function, along with an efficient projected gradient descent algorithm for model fitting. Simulation studies showcased that microTensor outperformed existing microbiome dimension reduction methods in terms of estimation variability, and had better power to differentiate subject-specific effects. We specifically note that our simulation
analyses were carefully designed such that a) the simulated microbial count observations have distribution characteristics similar to real data, and b) a real-world data based simulation was included, where the microbial observations were generated based on a real dataset, and low rank structures were not assumed a priori. Lastly, we performed two real-world dataset applications, where microTensor in practice detected effects on the human gut microbiome of diet, pharmaceuticals, and infant delivery modes.

There are a few directions for future improvements of our method. First, we observe that within subject correlations of microbiome samples in our model is fully specified through the mean structure model. That is, samples are correlated by sharing the same subject loading factor across different timepoints. This can be further generalized by allowing error terms to be also correlated across samples, similar as in a mixed effect generalized linear model. Inference for such models can be computationally difficult, especially given that our work is targeted for unsupervised dimension reduction, rather than supervised association testing applications. Identifiability considerations, as well as specialized optimization algorithms will likely be needed for such extensions. Second, while we adopted the canonical polyadic decomposition as the optimization algorithm, our model does not inherently require so. The CPD was used mainly for computational convenience, as well as based on its good empirical performance as reported in previous works (Mu, Hsu and Goldfarb (2015)). However, it is possible that a simultaneous optimization across all loading factors can help improve model fitting. For computation considerations, it is possible to adopt more complex optimization algorithms to improve model fitting.

Acknowledgment. This research was supported by NIH grants GM129781 and GM123056.

SUPPLEMENTARY MATERIALS. The online Supplemental Materials include additional simulation plots and real data analysis results. An R repository of the proposed method and all analyses performed is available at https://github.com/syma-research/microTensor.

REFERENCES


