Efficient and accurate inference of microbial trajectories from longitudinal count data

Tyler A. Joseph*[1], Amey P. Pasarkar[1], and Itsik Pe'er*[1,2,3]

[1]*Department of Computer Science, Columbia University, New York NY, USA*
[2]*Department of Systems Biology, Columbia University, New York NY, USA*
[3]*Data Science Institute, Columbia University, New York NY, USA*

Abstract

The recently completed second phase of the Human Microbiome Project has highlighted the relationship between dynamic changes in the microbiome and disease, motivating new microbiome study designs based on longitudinal sampling. Yet, analysis of such data is hindered by presence of technical noise, high dimensionality, and data sparsity. To address these challenges, we propose LUMINATE (LongitUdinal Microbiome INference And zero deTEction), a fast and accurate method for inferring relative abundances from noisy read count data. We demonstrate on synthetic data that LUMINATE is orders of magnitude faster than current approaches, with better or similar accuracy. This translates to feasibility of analyzing data at the requisite dimensionality for current studies. We further show that LUMINATE can accurately distinguish biological zeros, when a taxon is absent from the community, from technical zeros, when a taxon is below the detection threshold. We conclude by demonstrating the utility of LUMINATE for downstream analysis by using estimates of latent relative abundances to fit the parameters of a dynamical system, leading to more accurate predictions of community dynamics.
Code availability: `https://github.com/tyjo/luminate`

*tjoseph@cs.columbia.edu, itsik@cs.columbia.edu

1 Introduction

The human body is home to trillions of microbial cells that play an essential role in health and disease[5]. The gut microbiome, for instance, encodes over 3 million genes[20] responsible for a variety of normal physiological processes such as the regulation of immune response and breakdown of xenobiotics[6]. Disturbances in gut communities have been associated with several diseases, notably obesity[17] and colitis[18], and changes to the vaginal microbiome during pregnancy is associated with risk of preterm birth[7]. Consequently, investigating the human microbiome can provide insight into biological processes and the etiology of disease.

A major paradigm for microbiome studies design uses targeted amplicon sequencing of the 16S rRNA gene to produce read counts of each bacterial taxon in a sample[16]. Due to its low cost (compared to shotgun metagenomics), 16S rDNA sequencing is a valuable tool for generating coarse-grained profiles of microbial communities. Nonetheless, analysis of 16S datasets faces multiple domain-specific challenges. First, 16S datasets are inherently compositional[12]: they only contain information about the relative proportions of taxa in a sample. In addition, technical noise, such as uneven amplification during PCR, can produce read counts that differ substantially from the underlying community structure[16]. In particular, species near the detection threshold may fail to appear in a sample, necessitating a distinction between a biological zero — where a species is absent in the community — from a technical zero where it drops below the detection threshold[1]. Finally, the number of taxa and time points in a sample may be large, requiring methods that scale to high dimensional data.

Increasingly, study designs based on 16S rDNA sequencing have incorporated longitudinal sampling. This is exemplified by a major aim of the second phase of the Human Microbiome Project[19] being quantification of dynamic changes in the microbiome across disease-specific cohorts. Longitudinal sampling holds promise in elucidating causality between temporal changes in the microbiome and disease. It further provides a unique opportunity to address the statistical challenges of 16S sequencing by pooling information across longitudinal samples.

To this end, two recent methods have been proposed for analyzing noisy longitudinal count data: TGP-CODA[1] and MALLARDs[22]. TGP-CODA fits a Gaussian process model to longitudinal count data, providing estimates of denoised (latent) relative abundances and statistical correction for technical zeros. MALLARDs, dynamic linear models with multinomial observations, fit a state space model to count data to partition observed variation into biological and technical components. Both models highlight the importance of temporal modeling, and its utility in providing insight into microbial systems. However, efficient inference from time-series data is a challenging problem, and both methods have difficulty scaling with sample size and taxa.

1.1 Our contribution

We propose LUMINATE (LongitUdinal Microbiome INference And zero deTEction), an accurate and efficient method to infer relative abundances from microbial count data. Our contribution is two-fold. First, using variational inference we reformulate the problem of posterior inference in a state-space model as an optimization problem with special structure. Second, we propose a novel approach to differentiate between biological zeros and technical zeros.

We demonstrate on synthetic data that LUMINATE accurately reconstructs community trajectories orders of magnitude faster than current approaches. We further demonstrate LUMINATE's ability to accurately distinguish biological zeros from technical zeros. Finally, we demonstrate the utility of LUMINATE by using estimated relative abundances to infer the parameters of a dynamical system, leading to more accurate predictions of community trajectories.

2 Methods

2.1 Probabilistic Model of Latent Variables

Methods for inference from time-series data are often formulated using state-space models. State-space models describe latent dynamics as a sequence of time-indexed random vectors, $\boldsymbol{x}_t$, where $\boldsymbol{x}_t$ is dependent on time points in the past. Information about the hidden state of the system is obtained through noisy observation of each time point $\boldsymbol{y}_t$. Such models are well suited for describing microbial dynamics: $\boldsymbol{x}_t$ contain information about the true — hidden — relative abundances, while $\boldsymbol{y}_t$ are noisy sequencing reads. Furthermore, state-space models provide a flexible framework for more sophisticated modeling that better captures the data generating process. We include two additional variables important for modeling microbial count data: $\boldsymbol{w}_t$ which describes extinction and recolonization of taxa, and $\boldsymbol{z}_t$ which incorporates an additional layer of sequencing noise(Figure 1).

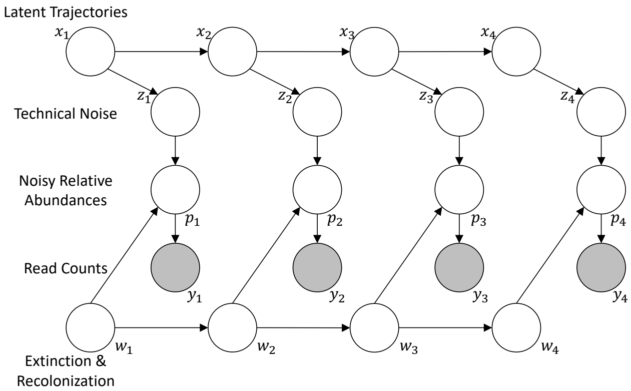

Figure 1: **Graphical model for 4 time points.** Sequencing counts $\boldsymbol{y}_t$ are determined by noisy relative abundances $\boldsymbol{p}_t$, which themselves are determined by the taxa alive at time t, $\boldsymbol{w}_t$, and noisy realizations $\boldsymbol{z}_t$ of the true community state $\boldsymbol{x}_t$.

Specifically, our model is as follows. Suppose we have a sample with T observed time points. Let $\boldsymbol{y}_t \in \mathbb{N}_0^D$ be the sequencing reads among D taxa at time t, and let $\boldsymbol{x}^t \in \mathbb{R}^{D-1}$ be the additive log ratio of the relative abundances of those taxa (the natural parameters of the multinomial distribution). The time between observations $t-1$ and t is denoted Δ_t. Further, let $\boldsymbol{z}_t \in \mathbb{R}^{D-1}$ be variables that represent noisy realizations of $\boldsymbol{x}_t$, and let $\boldsymbol{w}_t = (w_t^1, w_t^2, ..., w_t^D) \in \{0,1\}^D$ be indicator variables denoting which taxa are alive at time point t (i.e. $w_t^d = 1$ if taxa d is alive at time t, 0 otherwise). Our model is given by:

$$
\begin{aligned}
p(w_1^d = 1) &= \pi_i && d = 1...D \\
p(w_t^d = j | w_{t-1}^d = i) &= A_{ij}^d && d = 1...D,\ t = 1...T \\
p(\boldsymbol{x}_1) &= \mathcal{N}(\boldsymbol{x}_1 | 0, Q_0) && \\
p(\boldsymbol{x}_t) &= \mathcal{N}(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}, \Delta_t Q) && t = 1...T \\
p(z_t^d | x_t^d, w_t^d) &= \left(\mathcal{N}(z_t^d | x_t^d, r^d)\right)^{w_t^d} && t = 1...T \\
\boldsymbol{p}_t &\equiv \frac{1}{w_t^D + \sum_{d=1}^{D-1} w_t^d e^{z_t^d}} \left(w_t^1 e^{z_t^1}, ..., w_t^{D-1} e^{z_t^{D-1}}, w_t^D\right) && t = 1...T \\
p(\boldsymbol{y}_t | \boldsymbol{z}_t, \boldsymbol{w}_t) &= \text{Multinomial}(\boldsymbol{y}_t | N_t, \boldsymbol{p}_t) && t = 1...T
\end{aligned}
$$

The $\boldsymbol{z}_t$ describe additional sequencing noise not captured by the multinomial distribution. The multinomial distribution makes a strong assumption that the technical variance is purely due to otherwise uniform statistical sampling. The $\boldsymbol{w}_{1:T}^d$ constitute a hidden Markov model with transition probabilities $\{A_{ij}^d\}_{i,j \in \{0,1\}}$ describing the extinction and reintroduction of certain taxa, distinguishing biological from technical zeros. Setting $w_t^d = 0$ removes the contribution of z_t^d from the likelihood, and zeros out the relevant proportions in the multinomial counts. Conceptually the $\boldsymbol{w}_{1:T}$ approximate extinction and recolonization events by making them orthogonal to the state of the system. Finally, the $\boldsymbol{x}_t$ serve as a prior over the space of dynamics. The change in the system between time points depends on the covariance between ratios of taxa Q, and the time between observations Δ_t. By learning the posterior $\boldsymbol{x}_{1:T}|\boldsymbol{y}_{1:T}$, we can estimate relative abundances from sequencing counts through $\boldsymbol{x}_{1:T}$.

The covariance Q implicitly makes the assumptions that trajectories are smooth in time. However external perturbations such as antibiotics can rapidly induce changes in the community. We model these changes by introducing a perturbation covariance Q_p that replaces Q for time points with known (i.e. provided as input) perturbations.

Our model is conceptually similar to TGP-CODA[1] and MALLARDs[22]. Both models introduce variables analogous to $\boldsymbol{z}_t$ for technical noise, but take different approaches to modeling dynamics that come with increased computational cost. MALLARDs use a similar state-space model (that describes dynamics under a phylogenetically motivated log-ratio transformation). However, MALLARDs require evenly spaced time points — each time point occurs after a fixed interval of time. After specifying a unit of time, time points without observations are integrated out computationally using a Kalman filtering/smoothing approach. Additionally, MALLARDs do not incorporate terms for biological zeros as we do here. TGP-CODA, in contrast, incorporates additional variables for technical zeros similar to $\boldsymbol{w}_t$, but not true zeros which we claim the $\boldsymbol{w}_t$ represent. Furthermore, TGP-CODA learns a state-space covariance matrix using a Gaussian process model. This increased flexibility comes at a considerable computational burden.

2.2 Inference

Our main contribution is the demonstration that inference under such state-space models can be performed quickly using variational inference without loss of accuracy. By inference, we mean two things: posterior inference where the goal is to compute the posterior $p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}, \boldsymbol{w}_{1:T}|\boldsymbol{y}_{1:T})$, and parameter inference for the model parameters $A^{1:D}$, Q, and $r^{1:D-1}$. Variational inference transforms both inference problems to an optimization problem by approximating the true posterior $p_\theta(\cdot|\boldsymbol{y})$ with model parameters θ by a variational posterior $q_\nu(\cdot|\boldsymbol{y})$ with variational parameters ν. The parameters (θ, ν) are optimized to minimize the Kullback-Leibler divergence, or equivalently maximize the "evidence lower-bound", between the true and approximate posterior. The variational objective function is

$$\mathcal{L}(\boldsymbol{y}_{1:T}, \theta, \nu) = \mathbb{E}_q[\log p(\boldsymbol{w}_{1:T}, \boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}, \boldsymbol{y}_{1:T})] - \mathbb{E}_q[\log q(\boldsymbol{w}_{1:T}, \boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T})]$$

The main challenge in designing an inference algorithm for variational inference is choosing a form for q that is capable of closely approximating the true posterior while maintaining the ability to compute the expectations in $\mathcal{L}$ (while black-box approaches exist where the expectations in $\mathcal{L}$ are not explicitly computed, a closed form inference procedure is more desirable). Assuming a particular factorization of q and optimizing parameters using coordinate ascent, it is sometimes possible to compute an optimal parametric form for q for that also gives the optimal ν (see Blei *et al.*[3] for a derivation).

A common choice of factorization is to partition model variables into independent subsets

$$q(\boldsymbol{w}_{1:T}, \boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}) = \left[\prod_{d=1}^{D} q(\boldsymbol{w}^d_{1:T})\right] q(\boldsymbol{x}_{1:T})q(\boldsymbol{z}_{1:T}).$$

For this choice of factorization, the optimal $q(\boldsymbol{w}^d_{1:T})$ and $q(\boldsymbol{x}_{1:T})$ can be computed in closed form using block coordinate ascent (which we will show), while we need to make a choice for the parametric form of $q(\boldsymbol{z}_{1:T})$. A sensible choice for $q(z^d_t) = \mathcal{N}(z^t_d|\eta^d_t, \gamma^d)$. The η^d_t and γ^d are variational parameters that are optimized with respect to $\mathcal{L}$. The joint distribution across $\boldsymbol{z}_{1:T}$ is

$$q(\boldsymbol{z}_{1:T}) = \mathcal{N}(\boldsymbol{z}_{1:T}|\boldsymbol{\eta}_{1:T}, \Gamma)$$

where Γ is a diagonal covariance matrix with entries in $\{\gamma^1, ..., \gamma^{D-1}\}$. Given this choice of $q(\boldsymbol{z}_{1:T})$ the optimal choice of q's for $q(\boldsymbol{x}_{1:T})$ and $q(\boldsymbol{w}^d_{1:T})$ are given by[3]

$$\begin{aligned}
q(\boldsymbol{x}_{1:T}) &\propto \exp\left\{\mathbb{E}_{-\boldsymbol{x}_{1:T}}\left[\log p(\boldsymbol{w}_{1:T}, \boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}, \boldsymbol{y}_{1:T})\right]\right\} \\
&\propto \exp\left\{\mathbb{E}_{-\boldsymbol{x}_{1:T}}\left[\log p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T})\right]\right\} \\
q(\boldsymbol{w}^d_{1:T}) &\propto \exp\left\{\mathbb{E}_{-w^i_{1:T}}\left[\log p(\boldsymbol{w}_{1:T}, \boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}, \boldsymbol{y}_{1:T})\right]\right\} \\
&\propto p(w^i_{1:T}) \exp\left\{\mathbb{E}_{-w^i_{1:T}}\left[\log p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}, \boldsymbol{y}_{1:T}|\boldsymbol{w}_{1:T})\right]\right\}
\end{aligned}$$

where the expectations are computed with respect to all q except for the variable of interest. We devote the remainder of this section to demonstrating that these can be computed efficiently in closed form.

First, the joint distribution of $p(\boldsymbol{x}_{1:T}) = \mathcal{N}(0, \Lambda^{-1})$ is Gaussian with precision matrix Λ that is block tridiagonal. The simplest way to see this is to note that a Gaussian density is equivalent to its Laplace approximation. Hence, Λ is given by

$$\begin{aligned}
\Lambda &= \begin{bmatrix}
\Lambda_{1,1} & \Lambda_{1,2} & & & \\
\Lambda^T_{1,2} & \Lambda_{2,2} & \Lambda_{2,t} & & \\
 & \ddots & \ddots & \ddots & \\
 & & \Lambda^T_{T-2,T-1} & \Lambda_{T-1,T-1} & \Lambda_{T-1,T} \\
 & & & \Lambda^T_{T-1,T} & \Lambda_{T,T}
\end{bmatrix} \\
-\Lambda_{t,t} &= \frac{\partial^2}{\partial \boldsymbol{x}_t^2}\left[\log p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) + \log p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})\right] \\
-\Lambda_{t,t+1} &= \frac{\partial^2}{\partial \boldsymbol{x}_{t+1}\boldsymbol{x}_t}\left[\log p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) + \log p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})\right]
\end{aligned}$$

Simplifying $q(\boldsymbol{x}_{1:T}) \propto \exp\left\{\mathbb{E}_{-\boldsymbol{x}_{1:T}}\left[\log p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T})\right]\right\}$ leaves us with $q(\boldsymbol{x}_{1:T}) = \mathcal{N}(\boldsymbol{x}_{1:T}|\boldsymbol{\mu}_{1:T}, \Sigma)$ where Σ and $\boldsymbol{\mu}_{1:T}$ are given by

$$\Sigma = (\Lambda^{-1} + \Gamma^{-1})^{-1} \tag{1}$$

$$\Gamma^{-1}\boldsymbol{\eta}_{1:T} = \Sigma^{-1}\boldsymbol{\mu}_{1:T} \tag{2}$$

Notably, if we're only interested the posterior means $\boldsymbol{\mu}_{1:T}$, we never need to explicitly compute the entire posterior covariance Σ. Σ^{-1} is block tridiagonal, which means its inverse can be computed in $\mathcal{O}(TD^3)$ time instead of $\mathcal{O}(T^3D^3)$ time[13]. Furthermore, the solution for $\boldsymbol{\mu}_{1:T}$ only relies on the diagonal blocks of Σ^{-1} and an intermediate computation from the inverse. Consequentially, $\boldsymbol{\mu}_{1:T}$ can be computed in $\mathcal{O}(TD^2)$ after the inverse is computed, instead of $\mathcal{O}(T^2D^2)$.

Simplifying the expression for $q(\boldsymbol{w}^d_{1:T})$, reveals that the optimal $q(\boldsymbol{w}^d_{1:T})$ is given by

$$q(w^i_{1:T}) \propto p(w^i_{1:T}) \exp\left\{\mathbb{E}_{-w^i_{1:T}}\left[\log p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}, \boldsymbol{y}_{1:T}|\boldsymbol{w}_{1:T})\right]\right\}$$

This is precisely the posterior under a hidden Markov model with (now fixed) observations given by the exponential term. Moreover, the only terms we need to compute $\mathcal{L}$ are $q(w^i_t)$ and $q(w^i_t, w^i_{t-1})$, which can be computed in $\mathcal{O}(4T)$ time using the standard forward-backward equations for hidden Markov models (HMMs)[2].

Finally, the update for the parameters of $q(\boldsymbol{z}_{1:T})$ cannot be computed in closed form due to the structure of the problem. We instead rely on a conjugate gradient algorithm to optimize $\boldsymbol{\eta}_{1:T}$ (since $\boldsymbol{\eta}_{1:T}$ does not rely on the variance terms γ^d we choose not to optimize γ^d).

The only remaining difficulty is computing

$$\begin{aligned}\mathbb{E}_q[\log p(y_t|z_t,w_t)] &= \sum_{d=1}^{D-1} y_t^d\,\mathbb{E}_q[w_t^d]\,\mathbb{E}_q[z_t^d] - \mathbb{E}_q\left[N_t \log\left(w_t^D + \sum_{d=1}^{D-1} w_t^d e^{z_t^d}\right)\right] + \text{const}\\ &\geq \sum_{d=1}^{D-1} y_t^d\,\mathbb{E}_q[w_t^d]\,\mathbb{E}_q[z_t^d] - N_t \log\left(\mathbb{E}_q[w_t^D] + \sum_{d=1}^{D-1} \mathbb{E}_q[w_t^d]\,\mathbb{E}_q[e^{z_t^d}]\right) + \text{const}\end{aligned}$$

in $\mathcal{L}$. This lower bound on $\mathbb{E}_q[\log p(y_t|z_t,w_t)]$ bounds the objective $\mathcal{L}$ by below, which we note maintains a valid variational inference algorithm.

Once q has been formulated, optimizing model parameters $A^{1:D}, Q, r^{1:D-1}$ are straightforward. The expectations in $\mathcal{L}$ can all be computed (using the lower bound above), and taking the gradient with respect to each parameter and setting equal to zero obtain a closed form for each.

In summary, we have derived an inference algorithm for the model parameters and variational parameters of our model, where we can compute closed form block coordinate ascent updates for all but one set of parameters. Moreover, we can compute such updates efficiently by exploiting the special structure of the covariance of the state-space. Thus, we are left with the following algorithm.

```
Data: Sequencing counts $\boldsymbol{y}_{1:T}$
while $A^{1:D}, r^{1:D-1}, Q, \boldsymbol{\mu}_{1:T}, \boldsymbol{\eta}_{1:T}$ have not converged do
    Update $q(\boldsymbol{x}_{1:T})$ using equations 1 & 2;
    Update $q(w_t^d)$ and $q(w_t^d, w_{t-1}^d)$ using the forward-backward equations for HMMs;
    Update $\boldsymbol{\eta}_{1:T}$ using a conjugate gradient algorithm;
    Update model parameters $A^{1:D}, Q, r^{1:D-1}$ (all in closed form);
end
```

Algorithm 1: LUMINATE's inference algorithm

2.3 Simulation evaluation

We designed simulations to evaluate our model's ability to infer relative abundances from noisy sequencing data. To this end, we downloaded two dense longitudinal datasets of bacterial concentrations from Bucci *et al.*[4]: i) a dataset of 5 gnotobiotic mice colonized with a bacterial mixture of 16 species (the *C. diff* dataset), and ii) a dataset of 7 germ-free mice colonized with a mixture of 17 Clostridia strains (the *Diet* dataset). The *C. diff* dataset mice were subject to a *C. difficile* challenge after 28 days (average 26 observed time points observed over 56 days). The *Diet* dataset mice were fed a high-fiber diet for 5 weeks, switched to a low fiber diet for 2 weeks, then returned to the high-fiber diet for 2 weeks (average 47.14 observed time points across 65 days). We used these datasets to learn the parameters of a generalized Lotka-Volterra model (gLV)[26]. We chose to simulate trajectories using gLV because gLV has been shown to accurately describe microbial dynamics in some cases, in particular on the datasets we used to generate model parameters (see

Stein et al.[26] or Bucci et al.[4]).

$$\frac{d}{dt}\log x_i(t) = g_i + \sum_{j=1}^{D} A_{ij}x_j(t) + \sum_{p=1}^{P} B_{ip}u_p(t) \tag{3}$$

The $x_i(t)$ denote the concentration of bacteria i at time t, and the $u_p(t)$ denote external perturbations (such as introduction of *C. difficle* and change in diet). The parameters g_i, A_{ij}, and B_{ip} describe growth rates, interactions, and external effects respectively. We fit equation (3) by discretizing it and performing least squares with elastic net regularization, similar to Stein *et al.*[26].

Once we learned the model parameters for each dataset, we then forward simulated trajectories for each dataset given initial conditions of each mouse using the Runge-Kutta 5(4) method of numerical integration as implemented in `RK45` from `SciPy`[14]. This generated evenly spaced time points whose number corresponded to the number of observed time points of each mouse. We qualitatively inspected the simulated trajectories to ensure they matched the ground truth dynamics in the original data.

We simulated sequencing counts on top of each ground truth trajectory under varying levels of sequencing noise, following the framework of Silverman et al.[22]. Briefly, given temporal covariance Q and noise covariance R, they defined a signal-to-noise ratio as the total variance of Q over the total variance of R

$$\text{SNR} = \frac{\text{Tr}(Q)}{\text{Tr}(R)}$$

We computed the SNR under the additive log-ratio transformation: $\text{alr}(x_i(t)) = \log\left(x_i(t)/x_D(t)\right)$, using $\text{alr}(\boldsymbol{x}(t)) = (\text{alr}(x_1(t)), ..., \text{alr}(x_{D-1}(t))$ to compute the a diagonal covariance matrix Q of the state-space give by $\text{alr}(\boldsymbol{x}(t))$. The diagonal entries of Q measure how quickly each taxon $\text{alr}(x_i(t))$ changes over time. For fixed Q and fixed SNR, we set $R = \text{diag}\{r_1, ..., r_{D-1}\}$ where $r_i = \frac{q_i}{\text{SNR}}$. Thus, the sequencing noise was proportional to the variability of each taxa.

Finally, we simulated sequencing reads for each time point from the following model

$$\begin{aligned}
\boldsymbol{z}_t &\sim \mathcal{N}(\text{alr}(\boldsymbol{x}_t), R) \\
\log M_t &\sim \mathcal{N}(\log 10000, 0.5) \\
\log N_t &\sim \text{Poisson}(M_t) \\
\boldsymbol{p}_t &= \frac{1}{1 + \sum_{d=1}^{D-1} e^{z_t^d}} \left(e^{z_t^1}, ..., e^{z_t^{D-1}}, 1\right) \\
\boldsymbol{y}_t &\sim \text{Multinomial}(N_t, \boldsymbol{p}_t)
\end{aligned}$$

Intuitively, this means the average sequencing depth is approximately 10000 reads. The log-normal Poisson distribution on the number of sequencing reads N_t increases the variance in depth across samples, to better match the high variance of sequencing depth found in real data.

Importantly, all models we evaluated (see Section 2.5) make the same or more general assumptions about technical noise, and none assume gLV dynamics. Äijö et al.[1] assume a model equivalent to noise under the additive log-ratio transformation with additional noise from technical zeros, prior to observed sequencing counts. Silverman et al.[22] assume noisy realizations occur under the isometric log-ratio transformation (ilr). The ilr is a linear combination of the alr, and therefore simulating under the alr is equivalent to ilr under a linear transformation of the covariance matrix.

2.4 Biological zero detection simulations

To determine the ability of our model to detect biological zeros from technical zeros, we simulated 4 taxa across 30 days under gLV with carefully chosen parameters. We picked parameters such that one taxon would go extinct during the simulation, while forcing another taxon to remain near the detection threshold. The remaining 2 taxa were abundant throughout the simulation. This resulted in an approximately 2-to-1 ratio of true zeros versus technical zeros. We trained our model across 10 datasets of 10 longitudinal samples each, and for each observed zero computed the posterior probability that it was a biological zero: $q(w_t^d = 0)$.

2.5 Model comparison

We downloaded the code for TGP-CODA from GitHub[11]. As TGP-CODA only runs on a single sample at once, we ran it on each sample in each dataset individually using the default parameters, then combined the results. We estimated latent relative abundances by taking the mean of the posterior samples of variables Θ_G computing using the No-U-Turn Sampler (NUTS) in `PyStan`[24].

We downloaded the code for the MALLARD model from GitHub[10] , and extracted the code that performed posterior inference under their model in `RStan`[25]. Because the MALLARD implementation is not a complete software package, we needed to perform two modifications to the code to run on our simulated data. First, we used the canonical basis instead of the phylogenetic basis for the isometric log-ratio transformation. This results in no loss of generality because it only affects the interpretation of the coordinates of the state-space. Second, we changed how samples for MCMC were initialized. The original implementation used `RStan`'s black box variational inference algorithm to compute initial samples before running the NUTS sampler. However, `RStan`'s black box variational inference can fail unexpectedly, so we resorted to initializing samples using `RStan`'s default initialization. We estimated relative abundances by transforming posterior samples of θ to relative abundances, then taking the mean.

2.6 Utility for downstream analysis

We used estimated relative abundances from LUMINATE to fit the parameters of "compositional" Lotka-Volterra (cLV)[15], a nonlinear dynamical system describing microbial relative abundances we recently proposed. cLV uses the following model to describe changes in relative abundances across D over time:

$$\frac{d}{dt}\log\left(\frac{\pi_i(t)}{\pi_D(t)}\right) = g_i + \sum_{j=1}^{D} A_{ij}\pi_j(t) + \sum_{p=1}^{P} B_{ip}u_p(t) \qquad \text{for } i = 1...D-1 \tag{4}$$

The $\pi_i(t)$ give the relative abundance of taxon i at time t, g_i its relative growth rate, A_{ij} the relative interactions between taxa, and B_{ip} the effect of external perturbations. We learned the parameters A_{ij}, B_{ip}, g_i by discretizing (4) and performing least squares with elastic net regularization, trained on the $\pi_i(t)$ estimated by LUMINATE. Regularization parameters were chosen using leave-one-out cross validation, picking parameters with the lowest prediction error from initial conditions.

We compared LUMINATE + cLV to two other time-series models: the sparse autoregressive model (sVAR) proposed by Gibbons et al.[8] and the ARIMA-Poisson model proposed by Ridenhour et al.[21]. We download sVAR from GitHub[9], and ARIMA-Poisson from the supplementary material in Ridenhour et al.[21]. We fit both models following the methods from each respective paper: ARIMA-Poisson was fit with 1 time lag, sVAR was fit with 3 time lags. We further rarefied OTU counts to 10,000 reads for sVAR following Gibbons et al.[8].

We compared model performance on two datasets by predicting trajectories from initial conditions on test data. The first dataset was the *C. diff* dataset described above (the *Diet* dataset included concentrations only). We also used a dataset of 6 white-throated woodrats fed oxalate

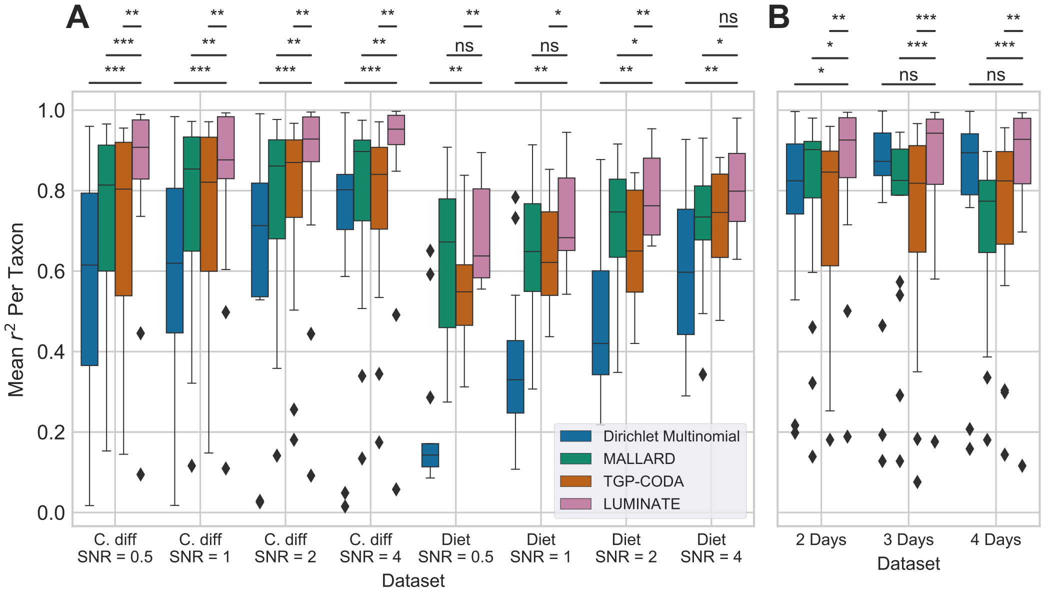

Figure 2: **LUMINATE accurately recapitulates relative abundance trajectories.** (A) Mean r^2 (y-axis) between ground truth and estimated relative abundances trajectories for each taxon. Equally spaced time points were simulated under generalized Lotka-Volterra with parameters learned from two real datasets (*C. diff* and *Diet*) with varying signal-to-noise ratio (SNR; x-axis). (b) Effect of sampling time on estimated trajectories under the *Diet* simulations with SNR=4. (Wilcoxon signed-rank test; ns: not significant; $*: p < 0.05$; $**: p < 0.01$; $***: p < 0.001$).

from Ridenhour et al.[21]. Using leave one-out cross validation, we predicted community trajectories on held out samples using model parameters learned on the remaining data.

3 Results

3.1 Simulations to assess the accuracy of LUMINATE

We first evaluated how well LUMINATE reconstructed (latent) community trajectories under varying amounts of sequencing noise. We generated ground truth trajectories by simulating data under generalized Lotka-Volterra (gLV) using parameters learned from real data (see Methods), then simulated noisy sequencing counts on top of each ground truth trajectory with varying signal-to-noise ratio and time between observations. We evaluated LUMINATE in comparison to three other models: i) a Dirichlet-Multinomial model (i.e a pseudocount model), ii) TGP-CODA[1], and iii) the specific MALLARD model from Silverman et al.[22]. Performance was compared by computing the mean r^2 between true and estimated trajectory for each taxon across longitudinal samples. This beneficially treats rare and common taxa on an equal scale.

Encouraging, LUMINATE had a significantly higher r^2 (p ¡ 0.05; Wilcoxon-signed rank test) than the Dirichlet-Multinomial model across all 8 of our simulations with evenly spaced time points (Figure 2A). We further observed significantly higher r^2 in 6 of 8 simulations when compared with the MALLARD model, and on 7 out of 8 simulations when compared with TGP-CODA. Importantly, LUMINATE performed no worse than the competing models across any of the simulations we investigated. Taken altogether, this suggests that LUMINATE is better recreating the latent community dynamics.

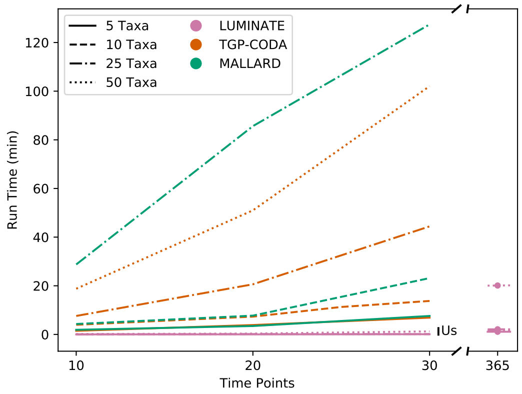

Figure 3: **LUMINATE is efficient.** Run time (measured as user time) in minutes (y-axis) for each model on a single longitudinal sample varying the number of time points (x-axis). Right: estimated run times for LUMINATE on 365 time points.

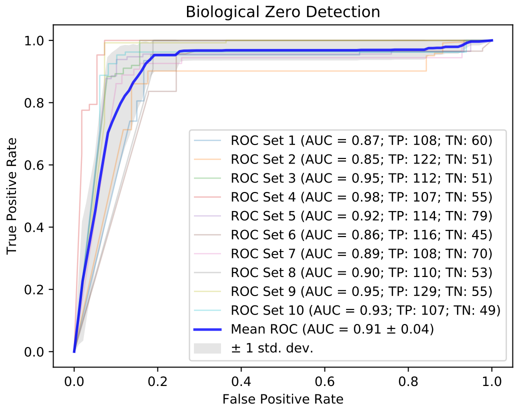

Figure 4: **LUMINATE accurately discriminates biological from technical zeros.** AUC-ROC curve using the posterior probability of a biological zero as a predictor for biological zeros on 10 simulated datasets. (TP: True Positives; TN: True Negatives)

Real microbiome datasets tend to be sparse in time. We therefore performed simulations to investigate sensitivity to technical frequency. We simulated data under learned parameters from *C. diff* data, and removed time points so that there was an observation every 2, 3, and 4 days on average. Notably, LUMINATE was robust to the sparser simulations (Figure 2B), outperforming TGP-CODA and the MALLARD model on all three simulations.

3.2 Simulations to assess the efficiency of LUMINATE

Both TGP-CODA and MALLARD models rely on Markov Chain Monte Carlo (MCMC) algorithms to compute posterior estimates of model variables. As MCMC can be computationally expensive, we wanted to evaluate how each model scales with increasing number of observed time points and taxa. We thus simulated a single longitudinal sample varying the number of time points and taxa.

Across all datasets, LUMINATE was faster then the other methods we investigated (Figure 3), sometimes by more than 2 orders of magnitude. LUMINATE ran in < 1.5 minutes on all datasets. In contrast, it to the MALLARD model 8.3 hours to run 50 taxa at 10 time points. On this same dataset it took TGP-CODA 18.28 minutes to run, but 1.7 hours to run on 50 taxa at 30 time points. In practice, this means that LUMINATE is the only method that can scale to datasets with multiple longitudinal samples and many observed taxa.

3.3 LUMINATE distinguishes biological zeros from technical zeros

We carefully designed simulations to test LUMINATE's ability to distinguish biological zeros — where a taxon is not presenting the community — from technical zeros, where it is below the detection threshold. Specifically, we simulated data where one taxon goes extinct over the course of the simulation, while another hovers near the detection threshold. For all zeros in the observed data, we computed the posterior probability of a biological zero, and evaluated performance by computing the area under the receiver operating characteristic (AUC-ROC). This measures the probability of the event that a biological zero receives a higher posterior probability than a technical zero, an indicator that the model differentiates between the two. We performed 10 replicates with 10 samples each to estimate confidence intervals for the AUC-ROC. Notably, the mean AUC was high across all replicates (Figure 4; mean = 0.91, std = 0.04), suggesting that our model accurately discriminates biological from technical zeros.

3.4 Utility for downstream analysis

We have demonstrated that LUMINATE accurately estimates relative abundances. These "denoised" estimates can be useful for downstream analysis of longitudinal data. One example is learning the parameters of a dynamical system. We recently proposed a nonlinear dynamical system called "compositional" Lotka-Volterra (cLV) that describes how relative abundances change over time. However, learning the parameters of cLV requires estimated relative abundances (as would any other dynamical system describing relative abundances). We thus asked if LUMINATE could be useful for fitting nonlinear models of microbial dynamics, and if such a nonlinear model could lead to better descriptions of the underlying dynamics.

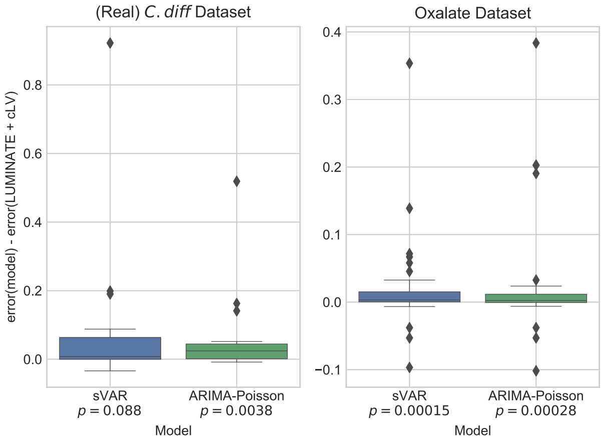

Figure 5: **LUMINATE improves downstream analysis.** Comparison of observed and predicted trajectories between cLV fit using LUMINATE and two other models. The y-axis displays the difference in error from each model (x-axis) to LUMINATE + cLV. Significance is computed using the Wilcoxon signed-rank test.

We fit cLV using LUMINATE's estimated abundances, and used cLV to forecast community trajectories from initial conditions. We compared forecast trajectories to two other models: ARIMA-Poisson[21] and (sVAR)[8]. For each model, we computed the average error between estimated and observed trajectories by taking Euclidean distance (the error) divided by the number of observed time points. We compared each model to LUMINATE + cLV by looking at the difference in error between the competing model and LUMINATE + cLV. This difference is expected to be symmetric around 0 if both models perform equally well (and greater than 0 if LUMINATE + cLV is performing better).

We observed that LUMINATE + cLV more accurately predicted trajectories than both sVAR and ARIMA-Poisson (Figure 5). For the sVAR model, this was significant in the Oxalate dataset. We observed a skew favoring LUMINATE + cLV on the *C. diff* dataset, but it did not reach the significance threshold, likely reflecting the smaller sample size (fewer taxa) in this dataset. In contrast, LUMINATE + cLV significantly outperformed ARIMA-Poisson on both datasets.

4 Discussion

Recent focus on dynamic changes in microbial communities has highlighted the importance of longitudinal modeling and data collection. Thus, there is an increasing need for methods for analyzing longitudinal data that are capable of scaling to large datasets spanning many taxa. With these goals in mind, we have proposed LUMINATE: a method for estimating relative abundances, and differentiating biological from technical zeros, in longitudinal datasets. We demonstrated that LUMINATE runs orders of magnitude faster than the current state of the art without loss of accuracy, can accurately detect biological zeros, and has utility as a preprocessing step for downstream analysis such as fitting the parameters of a dynamical system.

Though we emphasized variational inference as a tool to speed up computation, we note that this is not the only approach. In particular, Silverman et al.[23] propose an efficient algorithm for posterior inference in models they call marginally latent matrix-t processes, of which MALLARDs are a special case. However, there is currently no public implementation of MALLARDs in their framework. Still, MALLARDs do not distinguish biological from technical zeros, a major advantage of the present work.

There are several promising areas for future work. The true zero detection framework can be extended to include external perturbations, such as antibiotics, to assess how external factors affect risk of colonization by pathogenic bacteria. We can further expand our downstream analysis to learn biological interaction networks among taxa.

References

[1] T. Äijö, C. L. Müller, and R. Bonneau. Temporal probabilistic modeling of bacterial compositions derived from 16s rrna sequencing. *Bioinformatics*, 34(3):372–380, 2017.

[2] C. M. Bishop. *Pattern recognition and machine learning*. Springer Science+ Business Media, 2006.

[3] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[4] V. Bucci, B. Tzen, N. Li, M. Simmons, T. Tanoue, E. Bogart, L. Deng, V. Yeliseyev, M. L. Delaney, Q. Liu, et al. Mdsine: Microbial dynamical systems inference engine for microbiome time-series analyses. *Genome biology*, 17(1):121, 2016.

[5] I. Cho and M. J. Blaser. The human microbiome: at the interface of health and disease. *Nature Reviews Genetics*, 13(4):260, 2012.

[6] J. C. Clemente, L. K. Ursell, L. W. Parfrey, and R. Knight. The impact of the gut microbiota on human health: an integrative view. *Cell*, 148(6):1258–1270, 2012.

[7] D. B. DiGiulio, B. J. Callahan, P. J. McMurdie, E. K. Costello, D. J. Lyell, A. Robaczewska, C. L. Sun, D. S. Goltsman, R. J. Wong, G. Shaw, et al. Temporal and spatial variation of the human microbiota during pregnancy. *Proceedings of the National Academy of Sciences*, 112 (35):11060–11065, 2015.

[8] S. Gibbons, S. Kearney, C. Smillie, and E. Alm. Two dynamic regimes in the human gut microbiome. *PLoS computational biology*, 13(2):e1005364, 2017.

[9] GitHub. sVAR. `https://github.com/svazzole/sparsevar`, 2017.

[10] GitHub. MALLARD. `https://github.com/LAD-LAB/MALLARD-Paper-Code`, 2018.

[11] GitHub. TGP-CODA. `https://github.com/tare/GPMicrobiome`, 2018.

[12] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue. Microbiome datasets are compositional: and this is not optional. *Frontiers in microbiology*, 8:2224, 2017.

[13] J. Jain, S. Cauley, H. Li, C. Koh, and V. Balakrishnan. Numerically stable algorithms for inversion of block tridiagonal and banded matrices, submitted for consideration. *Numerical Linear Algebra Appl*, 2006.

[14] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. URL `http://www.scipy.org/`.

[15] T. Joseph, L. Shenhav, J. Xavier, E. Halperin, and I. Pe'er. Compositional Lotka-Volterra describes microbial dynamics in the simplex. *Under review*, 2019.

[16] J. Kuczynski, C. L. Lauber, W. A. Walters, L. W. Parfrey, J. C. Clemente, D. Gevers, and R. Knight. Experimental and analytical tools for studying the human microbiome. *Nature Reviews Genetics*, 13(1):47, 2012.

[17] R. E. Ley, P. J. Turnbaugh, S. Klein, and J. I. Gordon. Microbial ecology: human gut microbes associated with obesity. *Nature*, 444(7122):1022, 2006.

[18] C. Manichanh, L. Rigottier-Gois, E. Bonnaud, K. Gloux, E. Pelletier, L. Frangeul, R. Nalin, C. Jarrin, P. Chardon, P. Marteau, et al. Reduced diversity of faecal microbiota in crohn's disease revealed by a metagenomic approach. *Gut*, 55(2):205–211, 2006.

[19] L. M. Proctor, H. H. Creasy, J. M. Fettweis, J. Lloyd-Price, A. Mahurkar, W. Zhou, G. A. Buck, M. P. Snyder, J. F. Strauss, Weinstock, M. George, O. White, and T. I. H. M. Project. The integrative human microbiome project. *Nature*, 569(7758):641–648, 2019.

[20] J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59, 2010.

[21] B. J. Ridenhour, S. L. Brooker, J. E. Williams, J. T. Van Leuven, A. W. Miller, M. D. Dearing, and C. H. Remien. Modeling time-series data from microbial communities. *The ISME journal*, 11(11):2526, 2017.

[22] J. D. Silverman, H. K. Durand, R. J. Bloom, S. Mukherjee, and L. A. David. Dynamic linear models guide design and analysis of microbiota studies within artificial human guts. *Microbiome*, 6(1):202, 2018.

[23] J. D. Silverman, K. Roche, Z. C. Holmes, and L. A. David. Bayesian multinomial logistic normal models through marginally latent matrix-t processes. *arXiv preprint arXiv:1903.11695*, 2019.

[24] Stan Development Team. PyStan: the Python interface to Stan, 2019. URL `http://mc-stan.org/`. R package version 2.17.1.0.

[25] Stan Development Team. RStan: the R interface to Stan, 2019. URL `http://mc-stan.org/`. R package version 2.19.2.

[26] R. R. Stein, V. Bucci, N. C. Toussaint, C. G. Buffie, G. Rätsch, E. G. Pamer, C. Sander, and J. B. Xavier. Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLoS computational biology*, 9(12):e1003388, 2013.