Latent Variable Models: Maximum Likelihood Estimation and Microbiome Data Analysis

by

Chun Yu Hong

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor William Fithian, Co-chair
Professor Perry de Valpine, Co-chair
Professor Peng Ding
Professor Alan Hubbard

Spring 2020

Latent Variable Models: Maximum Likelihood Estimation and Microbiome Data Analysis

Abstract

Latent Variable Models: Maximum Likelihood Estimation and Microbiome Data Analysis

by

Chun Yu Hong

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor William Fithian, Co-chair

Professor Perry de Valpine, Co-chair

Data analysis often involves modeling complex relationships among many variables, some of which are unobserved. This type of analysis is usually tackled by latent variable models, which are graphical models consisting of both observed variables and latent variables. In this work, we delve into the computational aspect and the application aspect of latent variable models. On the computational side, we unify and extend stochastic gradient based maximum likelihood estimation methods for latent variable models under a framework called Hierarchical Model Stochastic Gradient Descent (HMSGD). Numerical studies have shown that certain extensions are more computationally efficient compared to the Monte Carlo Expectation Maximization (MCEM) algorithm. On the application side, we develop a nonparametric graphical model for microbiome data, and apply the framework to analyze the statistical properties of rarefaction, a popular normalization technique in microbiome data analysis. We show that rarefaction helps guarantee validity of permutation inference. We introduce the sample rarefaction efficiency index as a preliminary data-driven indicator of statistical efficiency of rarefied data compared to original data. Using the nonparametric graphical model, we propose a rarefaction-based nonparametric statistical testing procedure, the combined correlation permutation test, to assess whether library sizes are associated with microbial compositions conditioning on the grouping variable of interest. Case studies have shown that such associations are not uncommon in practice.

Contents

List of Figures

List of Tables

Acknowledgments

I am fortunate to have met many amazing people during my six-year PhD journey. To begin, I am extremely grateful to have William (Will) Fithian and Perry de Valpine as my advisors. Not only I have learned a great deal about statistics and ecology from our weekly meetings, but Will and Perry also have provided much guidance and support for me to grow as a researcher. I cannot express enough my gratitude for their patience going through many unexpected turns in my project directions. Will and Perry are supportive of my decision to working in the industry after graduation and they provide me valuable career advice.

I thank Xin Guo from the Industrial Engineering and Operations Research (IEOR) department for supervising me on my very first research project during my PhD on Bregman divergences, as well as a closely-related project on generative adversarial networks with her students Nan Yang and Tianyi Lin. I also thank Sara Stoudt, one of my fellow peers in the Statistics department, for collaborating with me on a project on maximum likelihood estimation for latent variable models, arise from a class project for Bayesian statistics taught by Chris Paciorek. This work eventually turns into Chapter 2 of my dissertation. I am also very grateful to work with Ulas Karaoz from Lawrence National Berkeley Laboratory (LNBL) on developing statistical methods for microbiome data analysis. Ulas had made many trips from LNBL to Evans to meet with me and he has been providing the much needed biological context for my research. I thank Alan Hubbard, Peng Ding, Perry de Valpine, and Will Fithian for being on my dissertation committee. Finally, I thank Peng Ding for introducing me to the world of causal inference.

During my time at UC Berkeley I enjoyed teaching very much and I had been a graduate student instructor (GSI) in almost every semester. I had great pleasure of working with, in chronological order, Ingileif B. Hallgrímsdottír, Hank Ibser, Shobhana Murali Stoyanov, Michael I. Jordan, Jarrod Millman, Ani Adhikari, David R. Brillinger, Gaston Sanchez, Merle Behr, Jared Fisher, and Adityanand Guntuboyina. Many thanks to the Statistics department for providing me the opportunity to be an unofficial instructor for STAT 153 in Spring 2018; this was my very first experience teaching as a lecturer during a regular school semester. I would also thank La Shana Polaris for helping me navigate through the logistics in the PhD program, as well as all the other wonderful staff members in the Statistics department for making my PhD journey a smoother ride.

I very much appreciate my officemates Billy Fang and Chelsea Zhang, as well as my best friends from my undergraduate studies Arthur Wu, Brian Shao, and Dulce Gomez, for listening to many of the ups-and-downs throughout my PhD. I would also like to extend my gratitude to Jason Wu, Zsolt Bartha, Satyaki Mukherjee, Aaron Schild, and Xiang Cheng for many board games nights and restaurant trips whenever I have to take a small break from my academics. I am greatly appreciative of people I met in Chinese Student Association at UC Berkeley in my final year. Their support and encouragement had made the final stretch of my PhD journey much more delightful.

Last but not least, my PhD journey would not have been possible without the invaluable support and nurturing of my parents.

Chapter 1

Introduction

1.1 Latent variable models

Graphical models are increasingly popular in modern data analysis because of their ability to describe complex relationships among many variables. Often we would like to relate observed variables to the unobserved variables. These unobserved variables are called *latent variables*. Graphical models that include latent variables are called *latent variable models*.

As a concrete application example, in ecology, researchers are often interested in modeling species community data to study species distributions and abundances. Often many species are of interest, so species-to-species associations are high-dimensional and these associations are hard to estimate directly without imposing any assumptions or structures. One approach is to impose a low-dimensional structure via the use of *latent factors*, assuming that the intricate associations can be explained by a few unobserved variables. This idea is used for example as part of the modeling specification in a flexible framework called Hierarchical Modeling of Species Communities (HMSC, [67]).

1.2 Parameter estimation

Broadly speaking, there are two approaches to parameter estimation for latent variable models: the *frequentist* approach and the *Bayesian* approach. There has been ongoing debate on the philosophy of these two approaches, but for our discussion we focus on the computational aspects.

In the frequentist approach, parameters are viewed as fixed, unknown values. A common frequentist estimation method is *maximum likelihood estimation*. In maximum likelihood estimation, parameter estimates are computed by finding the maximizer of the likelihood function, which is the joint density of the observed variables with parameters viewed as the function inputs. If the likelihood function has a closed-form expression, the problem of computing the maximum likelihood estimates is reduced to an optimization problem, which can be tackled by running an optimization algorithm such as Newton-Raphson or gradient descent, assuming a unique maximizer exists and the likelihood function is well-behaved. For latent variable models, the likelihood function typically does not have a closed-form expression, since it is often in terms of an intractable integral over the latent variables. A well-established general-purpose maximum likelihood estimation method for such models is the Monte Carlo Expectation Maximization (MCEM, [9, 24]) algorithm[1]. However, the runtime for MCEM can be quite long because of large sample sizes are typically required for proper approximation of the expectation in the E-step of the algorithm. Furthermore, while there are specialized software packages for running MCEM for specific models (for example, generalized linear mixed models), users often have to implement the details of the MCEM algorithms on their own if their models are more customized than the options availabe in the packages. There are very few general-purpose MCEM implementations available until recently (for example, in NIMBLE [65]).

In the Bayesian approach, parameters are viewed as random variables. Inference on parameters is done by studying the posterior distribution, the conditional distribution of the parameters given the data. Once the posterior distribution is obtained, point estimates of parameters can be obtained by computing the posterior mean or the posterior median. For many problems, when the analytical form of the posterior distribution is not available, one typically has to use a Markov Chain Monte Carlo (MCMC) method to sample from the posterior. JAGS (Just Another Gibbs Sampler, [74]), and more recently Stan [12] have made MCMC sampling from the posterior an easy-to-implement procedure once the model has been specified. Availability and user-friendliness of these software packages have arguably contributed to the popularity of the Bayesian approach to parameter estimation in latent variable models.

1.3 Microbiome data analysis

The advent in modern sequencing technologies has opened up many new exciting opportunities in probing the microbial worlds and facilitating new scientific discoveries. To compile a microbiome dataset from observations, researchers have to first perform an extraction procedure and polymerase chain reactions (PCR) to obtain 16S rRNA gene sequences. These sequences are then typically clustered at a certain similarity level to create groups of closely related microbes. One can think of these groups as an operational version of the idea of "species", more commonly called operational taxonomic units (OTUs) in the context of microbial biology. a A microbiome dataset is typically of the form of a count matrix, each entry representing the number of a particular OTU for a particlar observation. The dataset is commonly accompanied by information about the observations as well as phylogenetic relationships among the OTUs.

There are several statistical challenges when dealing with microbiome count data. First, microbiome datasets are high-dimensional. The number of features (OTUs), usually around 1000 to 10000, is typically much larger than the number of observations, usually around 10

[1] See Section 2.1 for other approaches to maximum likelihood estimation for latent variable models.

to 100, making comparisons among microbial communities difficult. Second, the data are typically over-dispersed and zero-inflated, rendering many traditional statistical approaches inappropriate. Third, numbers of sequences (known as *library sizes* or *read depths*) among the observations can vary greatly, so without proper normalization of observations comparisons may not be valid. Library size variations are often considered as artifacts of sequencing procedures.

One popular normalization strategy is rarefaction. The basic idea of rarefaction is to subsample each observation to the a pre-specified depth via sampling without replacement, and to discard all the observations with library sizes below the pre-specified depth. Previously, performance studies of rarefaction have been largely limited to simulation studies under specific parametric models. A non-parametric latent variable model on microbiome data could be very helpful for formalizing statistical properties of rarefaction.

1.4 Overview

The rest of the dissertation is organized as follows. In Chapter 2, we develop Hierarchical Model Stochastic Gradient Descent (HMSGD), a framework that unifies and extends stochastic gradient based methods for maximum likelihood estimation of latent variable models. Some of these extensions can arrive at the maximum likelihood estimates much faster than MCEM. In Chapter 3, we devise a non-parametric graphical model for microbiome data and formally study the statistical trade-offs of a popular normalization procedure called rarefaction. In Chapter 4, we develop a correlation-based permutation test for testing the association between the library size and the microbial composition and apply this test for various real-world datasets. Chapter 5 outlines a few future directions along the lines of research presented in this dissertation.

Chapter 2

Fast maximum likelihood estimation for general hierarchical models

This chapter is based on joint work with Sara Stoudt and Perry de Valpine.

2.1 Introduction

Hierarchical statistical models are widespread in the applied sciences because of their ability to capture complex relationships in data. In ecology, hierarchical models of species abundance through space and time have been applied to estimate species distributions and dynamics [81]. In political science, hierarchical models are used to estimate underlying preferences from data on policymaker decisions [43], while in epidemiology such models are used to estimate disease prevalence across space and time [56]. These models are difficult to estimate largely because the likelihood requires integration over the latent variables, which is typically a high-dimensional problem with no closed-form solution [18, 19].

Partly because of the difficulty of the likelihood integration problem, Bayesian analysis via computational tools such as Markov chain Monte Carlo (MCMC) has become the main practical path for analysis of many hierarchical models [20, 40]. However, many lines of statistical reasoning would be enabled by a similarly general computational approach for maximum likelihood estimation [91]. One might seek to apply likelihood ratio tests, model selection by AIC, goodness-of-fit as measured by maximum likelihood or other metrics, cross-validation, or other approaches. Although statisticians sometimes emphasize the philosophical incompatibility of Bayesian and frequentist results, practitioners are quite willing to study results from each side-by-side. Even when one seeks a Bayesian analysis, maximum likelihood results can provide a sanity check on the MCMC posterior and the influence of prior distribution assumptions. Statisticians have argued that the future will hold a combination of both Bayesian and frequentist methods [26], yet for general hierarchical models, practitioners are often limited to Bayesian results. Amid this rich space for statistical innovation, the need for improved MLE methods for general hierarchical models is vital.

A variety of methods have been proposed for MLE estimation of general hierarchical models, but none has gained the kind of general traction that MCMC has for Bayesian estimation. One set of methods are stochastic variants on the expectation maximization (EM) algorithm [24], such as Monte Carlo EM (MCEM, [98]), stochastic EM (SEM, [13, 14]), and stochastic approximation EM (SAEM, [11, 23, 53]). These suffer from the potentially slow convergence path of EM, and methods to ensure convergence involve costly increases in sample sizes to achieve smaller Monte Carlo variance as the algorithm proceeds [9]. Despite these issues, MCEM is one of the most widely applied methods because of its generality. A second approach, Monte Carlo Newton-Raphson (MCNR, [54]), also requires increasing Monte Carlo sample sizes to ensure convergence, although theory and application of this method appear less widespread. A third approach, data cloning [57] or State-Augmentation for Marginal Estimation (SAME, [44]), uses MCMC with many duplicate latent states, making the problem similar to MCMC but harder. A fourth approach, Monte Carlo Kernel Likelihood (MCKL, [90]), can require iterated application of MCMC.

Due to the various challenges in applying these methods, they are not used as widely as they might be. Specialized methods have been created for specific problems, such as stochastic approximation EM coupled with approximate Bayesian computation for state-space models (SAEM-ABC, [72]) and stochastic approximation EM with a Metropolis-Hastings sampling procedure that is based on a multidimensional Gaussian proposal for nonlinear mixed effects models (f-SAEM, [48]), but these fail to cover a wide range of latent variable model scenarios. Finally, we note that there has been interest in distributional approximation methods such as INLA [82] and variational Bayes [5, 95] for the related Maximum Posterior Estimation problem, but we focus on problems where the goal is exact MLE limited only by small Monte Carlo error.

We exploit a connection between the hierarchical model maximum likelihood problem and stochastic gradient descent methods in non-hierarchical models to obtain efficiency improvements that range from 1-2 orders of magnitude in three examples. This level of improvement has the potential to change statistical practice in analysis of hierarchical models by enabling more routine maximum likelihood estimation.

In typical stochastic gradient descent problems, one has a non-hierarchical model for "big data". For example, the loss function for neural network parameters is a sum of many terms, and the computation of its exact gradient to take an optimization step is costly [6, 7]. It turns out to be more efficient to calculate a gradient from a stochastic subset of the data at each iteration and to use a running average of steps to smooth over the stochasticity [7, 37, 38]. In essence, many fast, noisy steps converge more efficiently than few slow, deterministic steps.

In a hierarchical model, the relevant gradient is an expectation over latent states, often approximated by Monte Carlo. In this context, the large sample is not the data but rather the Monte Carlo sample of latent states given parameters and data. Existing approaches represent different ways of taking approximately deterministic steps at the cost of large Monte Carlo sample sizes. The connection to stochastic gradient descent methods suggests that many faster but noisier steps may work better. We take advantage of the fact that highly developed step-size schedules from the stochastic gradient descent literature are directly transferable to the hierarchical model maximum likelihood problem in ways that have not been done before. The general view of the problem also leads us to propose a new method based on a greedy line search in an approximate gradient direction, which is better than previous methods but not best in our computational experiments.

While the connection between maximum likelihood estimation of hierarchical models and stochastic gradient descent of non-hierarchical models has been recognized before in the computer science and machine learning literature [86, 87] and discussed in the context of Hidden Markov Models [10], we have not found significant crossover of these ideas to the applied statistics literature to achieve large efficiency gains as we do here. By placing methods in a common framework, we can see them as variants on how to iterate between sampling latent states given data and current parameters and making a step to update parameters. The methods differ in sample-size and step-size choices, which we show can be improved by drawing on advances in stochastic gradient descent methods. (Despite that the likelihood is being maximized, we stick with the established label "descent", viewing the negative log likelihood as a loss function to be minimized.) Specifically, we compare the fixed step-size schedule, the second-order based step-size schedule (Newton-Raphson), and the adaptive step-size schedule Adam [51]. We refer to this class of methods as Hierarchical Model Stochastic Gradient Descent (HMSGD). To the best of our knowledge, applying Adam step-size schedules to Monte Carlo maximum likelihood algorithms for general hierarchical models has not been studied, and it is these methods that yield the best and most stable performance.

The new step-size schedule we propose, called iterative 1D sampling, emerges naturally as a combination of the gradient-descent view of the problem and the Monte Carlo Kernel Likelihood idea. The essential idea is to draw MCMC samples of the latent states and parameters, but with parameters constrained to move in the direction of the approximated gradient at the values from the previous maximization. The new maximizer in this direction is then approximated using kernel density estimation (Figure 2.1). Section 2.3.3.3 describes this approach in detail.

In Section 2.2, we establish notation for a general hierarchical model. In Section 2.3, we place MCEM, MCNR, and Hierarchical Model Stochastic Gradient Descent in a common framework and introduce Adam as a viable step-size schedule. We also introduce the greedy stochastic line search method. In Section 2.4, we discuss computational considerations, and in Section 2.5 we present computational experiments from three examples. The examples include a Gamma-Poisson mixture model of pump failure times, a logistic GLMM with random intercepts for seed germination, and a logistic GLMM for salamander mating success with crossed random effects. Crossed random effects present a challenge to many numerical methods. Results show that Adam and the newly proposed iterative 1D sampling can achieve reasonably accurate estimates even with small MCMC sample sizes, leading to remarkable improvement in computational time. Section 2.6 provides discussion and directions for future work; Section 2.7 concludes the chapter.

We implement all of these methods in NIMBLE (Numerical Inference for statistical Models for Bayesian and Likelihood Estimation, [66, 92]. NIMBLE is an R package that allows for flexible hierarchical model specification and writing algorithms such as MCMC or the methods proposed here that can adapt to different model structures. The system is extensible and automatically generates model- and algorithm-specific C++ for fast execution.

2.2 Model Setup

Suppose we have n observations $y = (y_1, ..., y_n) \in \mathcal{Y} \subset \mathbb{R}^n$, drawn from probability distribution $p(y|\theta)$, where $\theta = (\theta_1, ..., \theta_D) \in \Theta \subset \mathbb{R}^D$ are the model parameters. We introduce the latent variables $x = (x_1, ...x_K) \in \mathcal{X} \subset \mathbb{R}^K$, which are considered unobserved random variables. The general latent variable model structure is as follows:

$$x|\theta \sim p(x|\theta) \tag{2.1}$$

$$y|x, \theta \sim p(y|x, \theta). \tag{2.2}$$

The (marginal) likelihood of θ is

$$L(\theta) := p(y|\theta) = \int p(y|x, \theta)p(x|\theta)dx \tag{2.3}$$

and our goal is to find

$$\hat{\theta}^{ML} := \arg\max_{\theta \in \Theta} \log p(y|\theta). \tag{2.4}$$

We are often interested in the scenario where the dimension of the latent variables is much larger than the dimension of the parameters; i.e. when $D \ll K$. When the dimension of the latent variables K is large, it is computationally infeasible to approximate the integral using a grid-based numerical integration. On the other hand, if D is small (say $D = 2$), it is tempting to use the naive Monte Carlo approximation

$$\frac{1}{S}\sum_{i=1}^{S} p(y|x^{(s)}, \theta)$$

based on $x^{(s)} \sim p(x|\theta)$ for a grid of values of θ and find the maximizer. However, this rarely works well due to high variance, since the $x^{(s)}$ are drawn from a distribution without information from the data y. To remedy the high variability, the sample size S has to be set at a large value, which in turn increases the computational cost dramatically.

2.3 A General Framework for Sampling-based Optimization Approaches

We begin by giving the general framework within which MCEM, MCNR, and gradient descent are special cases.

Given a current iterate $\theta^{(t)}$, each of the algorithms performs the following two steps:

1. **Sample Step:** Generate MCMC samples $x^{(t)} = (x^{(t),1}, x^{(t),2}, ...x^{(t),S})$ from $p(x|y, \theta^{(t)})$.

2. **Move Step:** Update $\theta^{(t+1)} = f(x^{(t)})$.

where the choice of f is different for each algorithm. The notation we use emphasizes the dependence of f on the MCMC sample $x^{(t)}$, but f can also depend on any quantities involved in the computation up to iteration t.

For the move step in most sampling-based approaches, gradient computation of the complete log-likelihood is required. This can be seen from from the Fisher's identity [27]:

$$\frac{d}{d\theta} \log p(y|\theta) = \mathbb{E}_{X \sim p(x|y,\theta)} \left[\frac{d}{d\theta} \log p(X, y|\theta) \right]. \tag{2.5}$$

The gradient computation inside the expectation can be done efficiently via an autodifferentiation package, and the expectation can be approximated via a Monte Carlo method.

When autodifferentiation packages are not available, one can use a finite-element approximation [3]:

$$\begin{aligned} \frac{d}{d\theta} \log p(y|\theta) &= \frac{1}{p(y|\theta)} \frac{d}{d\theta} p(y|\theta) \\ &\approx \left(\frac{\frac{p(y|\theta+\delta e_1)}{p(y|\theta)} - 1}{\delta}, \cdots, \frac{\frac{p(y|\theta+\delta e_D)}{p(y|\theta)} - 1}{\delta} \right), \end{aligned} \tag{2.6}$$

where e_i denotes the unit vector in the ith coordinate and δ denotes a very small value, say 10^{-4}. This suggests that the key to the approximation is to estimate the ratio $\frac{p(y|\theta+\delta e_i)}{p(y|\theta)}$. Following [33], note that

$$\frac{p(y|\psi)}{p(y|\theta)} = \frac{1}{p(y|\theta)} \int p(x|\psi) p(y|x, \psi) dx$$

and for any x,

$$\frac{1}{p(y|\theta)} = \frac{p(x|y, \theta)}{p(x|\theta) p(y|x, \theta)}$$

. Therefore,

$$\frac{p(y|\psi)}{p(y|\theta)} = \int \frac{p(x|\psi) p(y|x, \psi)}{p(x|\theta) p(y|x, \theta)} p(x|y, \theta) dx,$$

which can then be estimated by a standard Monte Carlo estimate using MCMC draws $x^1, ..., x^S$ from $p(x|y, \theta)$,

$$\frac{1}{S} \sum_{s=1}^{S} \frac{p(x^s|\psi) p(y|x^s, \psi)}{p(x^s|\theta) p(y|x^s, \theta)}.$$

Letting $\psi = \theta + \delta e_i$ for $i = 1, ..., D$ provides approximation of the likelihood ratios needed in the finite-element approximation of the gradient.

Higher-order derivatives can be computed in a similar fashion [10]. For example, the Hessian of the log-likelihood admits the following representation

$$\begin{aligned}\frac{d^2}{d\theta d\theta^T}\log p(y|\theta) &= \mathbb{E}_{X\sim p(x|y,\theta)}\left[\frac{d^2}{d\theta d\theta^T}\log p(X,y|\theta)\right]\\ &+\mathbb{E}_{X\sim p(x|y,\theta)}\left[\left(\frac{d}{d\theta}\log p(X,y|\theta)\right)\left(\frac{d}{d\theta}\log p(X,y|\theta)\right)^T\right]\\ &-\left(\frac{d}{d\theta}\log p(y|\theta)\right)\left(\frac{d}{d\theta}\log p(y|\theta)\right)^T,\end{aligned} \tag{2.7}$$

which is often known as Louis' identity [58]. This will prove useful in MCNR as discussed in Section 2.3.2. Finite-element approximations of higher-order derivatives require division by extremely small values and hence are numerically unstable. In a preliminary version of this work, we used finite-element approximation for Hessians in MCNR and found that the algorithm often diverged very quickly in our numerical studies. This issue is no longer present once automatic differentiation is used.

Define respectively $G_{MC}(\theta,x)$ and $H_{MC}(\theta,x)$ as the Monte Carlo approximation of (2.5) and (2.7) at θ based on a sample x. Now we present MCEM, MCNR, and Stochastic Gradient descent and show how they fit into this unifying framework. Each method uses the same sample step, but the function f for the move step varies.

2.3.1 Optimization as the Move Step

The MCEM algorithm [98] replaces the expectation in the E-step of the traditional EM algorithm [24] with a Monte Carlo approximation. The corresponding move step is

$$f(x^{(t)}) = \arg\max_{\theta}\frac{1}{S}\sum_{i=1}^{S}\log p(x^{(t),i},y|\theta). \tag{2.8}$$

2.3.2 Second Order Move Step

Monte Carlo Newton-Raphson (MCNR) uses Monte Carlo estimates of the gradient and the Hessian of $\log p(y|\theta)$ for Newton-Raphson updates [63, 54]. The corresponding move step is

$$f(x^{(t)}) = \theta^{(t)} - [H_{MC}(\theta^{(t)},x^{(t)})]^{-1}G_{MC}(\theta^{(t)},x^{(t)}). \tag{2.9}$$

2.3.3 First Order Move Step

The first-order move step is obtained by replacing the Hessian in (2.9) with a step-size choice α. To allow each component to have its own step-size, we consider α to be a D-dimensional vector, where D is the number of components in the parameter vector. Denoting the component-wise product (known as the Hadamard product) as $\odot$, the first order move step can be written as

$$f(x^{(t)}) = \theta^{(t)} - \alpha \odot G_{MC}(\theta^{(t)}, x^{(t)}). \tag{2.10}$$

We note that two common step-size choices are not useful for our problem. The first one is an inexact line search based on Wolfe conditions. Wolfe conditions guarantee sufficient improvement in the iterate and a decrease in the magnitude of the gradient [102]. Unfortunately, checking Wolfe conditions in our context is computationally costly, since it requires multiple evaluations of the marginal likelihood at each iteration. The second one is the Robbins-Monro step-size schedule [78], a popular step-size schedule for root finding, applications in regression [50], probability density estimation [49], and stochastic gradient methods in training neural networks. However, we found in our experiments that the Robbins-Monro step-size in our context leads to slow convergence compared to the alternative step-size schedules we considered and requires careful tuning. We therefore omit the Robbins-Monro results below.

We will now describe four ways to select the step-size: fixed step-size, Adam, and one-dimensional greedy line search.

2.3.3.1 Fixed step-size

The fixed step-size method suggests the use of a fixed learning rate for each component. Tuning the size for a particular problem can be tricky to automate, so we appeal to other choices of α that rely less on a user explicitly tuning the method in the following sub-sections.

2.3.3.2 Adam

Adam ([51]) uses bias-corrected moment estimates of gradients. Instead of using the estimated gradient $G_{MC}(\theta^{(t)}, x^{(t)})$ at the current iterate, the move is governed by an adjusted gradient. For $i = 1, ..., D$, define the running averages of first and second-order moment estimates of gradient:

$$m_i^{(t+1)} = \beta_1 m_i^{(t)} + (1-\beta_1)[G_{MC}(\theta^{(t)}, x^{(t)})]_i; \tag{2.11}$$

$$v_i^{(t+1)} = \beta_2 v_i^{(t)} + (1-\beta_2)([G_{MC}(\theta^{(t)}, x^{(t)})]_i)^2, \tag{2.12}$$

where β_1, β_2, α and ϵ are predetermined fixed scalars, and $m_i^{(0)}$ and $v_i^{(0)}$ are set to zero. Define bias-corrected first and second-order moment estimates:

$$\hat{m}_i^{(t+1)} = \frac{m_i^{(t+1)}}{1-\beta_1^{t+1}}; \tag{2.13}$$

$$\hat{v}_i^{(t+1)} = \frac{v_i^{(t+1)}}{1-\beta_2^{t+1}}. \tag{2.14}$$

The Adam update step is $f(x^{(t)}) = \theta^{(t)} - \alpha_{\text{adam}} \odot G^{(t)}_{MC,\text{adj}}$ with

$$\alpha_{\text{adam}} = \left[\frac{\alpha}{\sqrt{\hat{v}_1^{(t+1)}} + \epsilon}, ..., \frac{\alpha}{\sqrt{\hat{v}_D^{(t+1)}} + \epsilon} \right]$$

and $G^{(t)}_{MC,\text{adj}} = [\hat{m}_1^{(t)}, ..., \hat{m}_D^{(t)}]$.

In the context of stochastic gradient methods, convergence results are often stated in terms of bounds on the average regret, defined as

$$\frac{1}{T} \sum_{t=1}^{T} f_t(\theta_t) - \min_{\theta'} \frac{1}{T} [\sum_{t=1}^{T} f_t(\theta')$$

with f_t being a sequence of convex loss functions. [51] shows that Adam achieves an average regret bound of $O(1/\sqrt{T})$ under mild conditions, one of which is the convexity of the objective function. While in general a hierarchical model might not be globally convex, provided that the sample size (of y) is large enough, often the likelihood surface is locally similar to a Gaussian distribution near the optimum and hence locally convex.

2.3.3.3 One Dimensional Greedy Line Search

The idea of using an adaptive step-size has been explored in stochastic approximation and gradient methods [73, 107]. We introduce a novel adaptive step-size method called *1D greedy line search* that has its roots in Monte Carlo likelihood estimation by weighted posterior kernel densities (MCKL, [90]). The idea is to optimize the step-size at each step by solving the following optimization problem

$$c_t^{\max} = \arg\max_c p(y|\theta^{(t)} + cG_{MC}(\theta^{(t)}; x^{(t)}), \tag{2.15}$$

and then updating $\theta^{(t+1)} = f(x^{(t)})$, where

$$\alpha^{(t)}_{\text{1D}} = [c_t^{\max}, ..., c_t^{\max}]$$

and

$$f(x^{(t)}) = \theta^{(t)} + \alpha^{(t)}_{\text{1D}} \odot G_{MC}(\theta^{(t)}; x^{(t)}).$$

In essence we are always choosing the “best” step-size in the sense that we pick the one that provides the most progress. To approximately solve the optimization problem, given $(x^{(t)}, \theta^{(t)})$, we sample jointly in (x, γ) from

$$\tilde{p}(x, \gamma|y) \propto p(x, \theta^{(t)} + \gamma G_{MC}(\theta^{(t)}; x^{(t)})|y).$$

The samples of γ approximate a one-dimensional slice of the marginal distribution. We approximate the maximizer on the line using a kernel density estimate of the MCMC samples.

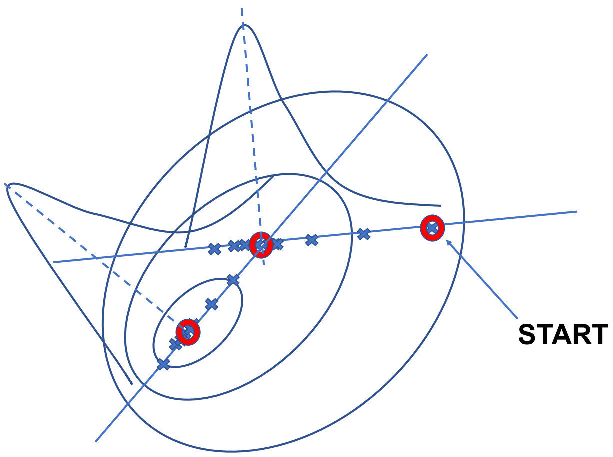


Figure 2.1: Visualization of the 1-dimensional sampling. The ellipses represent the contours of the likelihood surface. The blue crosses indicate the MCMC samples and the blue curves represent the density estimates. Each of the red circles indicates the parameter estimate at an iteration of the algorithms, which is computed as the mode of the estimated density.

The advantage of using a 1D line search is the potential of aggressive moves at the start of the algorithm. The downside of 1D line search is the computational cost incurred by additional MCMC sampling at each step. In addition, the number of MCMC samples needed for the greedy line search has to be reasonably large in order for the kernel density estimation (and hence the mode estimation) to be reliable.

Algorithm 1 Gradient descent via 1D Greedy Line Search

```
    • Input: $\theta^{(t)}$, $g^{(t)} := G_{MC}(\theta^{(t)}; x^{(t)})$.
Run an MCMC sampler to sample $(x^{(t)}, \gamma^{(t)})$ from $\tilde{p}(x, \gamma|y) \propto p(x, \theta^{(t)} + \gamma g^{(t)}|y)$.
Perform a 1D kernel density estimate for the sampled $\gamma^{(t)}$ and compute the mode $\hat{\gamma}^{(t)}$.
Set $\theta^{(t+1)} = \theta^{(t)} + \hat{\gamma}^{(t)} g^{(t)}$.
```

2.4 Computational Considerations

2.4.1 Burn-In and Warm Start

We set the burn-in to be half of the MCMC samples to be conservative (Section 6.5 of [8]). Since diagnostics require considerable amount of time to run and assess, it is more beneficial to run more samples (and conservatively remove the first half of the samples) rather than worry about tracking diagnostics throughout each step. We also implement a "warm start" where the last draws from the previous iteration's chain are the starting point for the next iteration. This idea is used in many contexts [22, 80, 97] and in our case should also help ensure that the burn-in period is adequate. We leverage the progress from the previous iteration, taking advantage of the proximity of the parameter estimates between two adjacent iterations.

2.4.2 Kernel and Bandwidth for the 1D Sampling Approach

For the kernel density estimation involved in the 1D sampling, we find that almost any reasonable choice of the kernel and the bandwidth yields similar final maximum likelihood estimates. For the numerical experiments, the kernel choice is defaulted to be Gaussian and the (optimal) bandwidth is computed based on the effective MCMC sample size instead of the nominal sample size, accounting for the fact that the usual optimal bandwidth is derived based on an independently identically distributed (i.i.d.) assumption [84].

2.4.3 MCMC sample sizes

Just as the number of MCMC samples that we use for the gradient estimation is a hyperparameter, analogous to the batch size in stochastic gradient descent, that can be tuned, so is the number of MCMC samples for 1D sampling. For the 1D sampling, a larger MCMC sample size is needed due to the slow mixing of joint sampling of the parameters and latent variables. In our experiments, the MCMC sample size choice[1] of 300 seems to work reasonably well. In one of our case studies, reasonably accurate performance can still be achieved with only a sample size of 20.

2.4.4 Convergence Criteria

For stochastic gradient descent and its variants, the convergence is often checked by predictive performance in machine learning applications. This is not appropriate in our MLE context since we are not solving a prediction problem. Within MCNR, [54] uses a formal hypothesis testing procedure where the variance of the updates are deduced based on a bootstrapping procedure. This is not applicable to our framework either since the sample sizes involved in our algorithms tend to be too small for proper variance estimation. Lastly, a less formal option for convergence check is to plot the trajectory of the iterates and see if the trajectory for each parameter roughly fluctuates around a particular number [9]. This requires users to study the trajectory plots, which is not ideal for an automated MLE algorithm.

Our approach is to quantify the fluctuation and flatness of the estimate trajectory, leading to the development of a two-step test for convergence. This approach is chosen to be ad hoc but fast. Once the two-step test is passed, the algorithm is terminated. For the first step, we use the Wald-Wolfowitz runs test [96] to determine whether the trajectory is close to being

[1] The number of samples used in the gradient approximation or the 1D sampling is only 150, since the first half of the samples are discarded. Similar comments also apply for a different choice of sample size.

monotonic, an indication of non-convergence. More specifically, at the tth iteration of the algorithm, check if the number of runs in $(\theta_j^{(t-w+1)}, \theta_j^{(t-w+2)}, ..., \theta_j^{(t)})$ are at least $r = 4$ for a block size $w = 20$. If this is not the case for every coordinate, continue the algorithm. If all the coordinates have runs of at least r, we proceed to the next test. The choices of r and w are somewhat arbitrary here and can be tuned. For the second step, we carry out a t-test to compare the average of iterates in the most recent 20 iterations with that in the preceding 20 iterations. If we fail to reject the null hypothesis at a certain significance level (say 30%), we terminate the algorithm and conclude convergence. Note that a larger significance level means we are being conservative, and therefore we will favor running more iterations.

We remark that for MCEM in our numerical experiments, we use the implementation in NIMBLE, which is based on [9]. The implementation gradually increases the MCMC sample sizes, unlike our proposed approach of fixing a small MCMC sample size. Since the gradient estimates are reliable for large MCMC sample size, this allows the simple convergence check: check whether the approximated gradient is within a certain tolerance.

2.5 Numerical Experiments

We experiment with the algorithms using three examples: a conjugate Gamma-Poisson hierarchical model (referred to as *pump*) and two GLMMs (referred to as *seeds* and *salamander*). For gradient and Hessian computations, we use automatic differentiation instead of finite-element approximation for better speed and higher accuracy. We run each algorithm for 300 iterations and report the execution times. In addition, if the algorithm passes the convergence test within 300 iterations, we report the convergence time and the number of iterations to convergence. To obtain the final estimates, we take the 20% trimmed mean of the last 20 iterates. The averaging is to smooth out any "bouncing around" the optimum towards the end of the path, and the trimming is to make the estimate more robust to occasional deviations on the path. Similar stabilizing approaches exist in the stochastic gradient descent literature such as taking the average of the last α proportion of iterates, called α-suffix averaging [75].

For every algorithm in each example, we report the CPU execution time (for 300 iterations), the CPU time to convergence (that is, how long it takes until the convergence test passes), the log-likelihood difference, and the mean-squared error (MSE) of the estimates compared to the benchmark estimates. Detailed numerical results can be found in Supplementary Materials.

2.5.1 Case study: pump (Gamma-Poisson hierarchical model)

The *pump* model [32] is a classic example from WinBUGS [62]. It is a conjugate Gamma-Poisson hierarchical model, so the marginal likelihood can be analytically derived. The MLE can be found by a standard deterministic optimization procedure[2]. The model specification

[2] We use `optim()` in R.

is as follows: for $i = 1, ..., N$,

$$\theta_i|\alpha, \beta \sim \text{Gamma}(\alpha, \beta), \tag{2.16}$$

$$\lambda_i = \theta_i t_i, \tag{2.17}$$

$$x_i|\lambda_i \sim \text{Poisson}(\lambda_i), \tag{2.18}$$

where x_i is the number of failures for pump i, θ_i is the failure rate for pump i, and t_i is the length of operation time for pump i. We treat $x_1, ..., x_N$ as the observed random variables, $t_1, ..., t_N$ as fixed constants and $\theta_1, ..., \theta_N$ as latent variables. The pump reliability dataset is originally from [31]. It consists of data about ten power plant pumps ($N = 10$).

To investigate the sensitivity of the algorithms to initial values, each of the algorithms is tested with two different starting points $(\alpha^{(0)}, \beta^{(0)}) = (10, 10)$ and $(\alpha^{(0)}, \beta^{(0)}) = (10, 2)$. We observe that all the algorithms except the smaller fixed step-size (0.005) are able to get close to the benchmark MLE. Figure 2.2 shows that the 1D sampling approach makes aggressive moves initially, so it gets close to the optimum in fewer iterations. The smaller fixed step-size (0.005) follows the shape of the likelihood surface more closely at the cost of many more steps. Adam's final iterations concentrate more tightly around the optimum. The methods are robust to both different initial values and different MCMC sample sizes. The latter robustness allows us to reduce the computational time by not relying on as many MCMC samples to assure good performance. We remark that MCEM takes far fewer iterations but a much longer computational time to reach the optimum.

2.5.2 Case study: seeds (logistic regression with random effects)

Our next example, *seeds*, is a logistic regression model with random effects. These types of models are common in social sciences and medicine as many longitudinal studies have a binary outcome [68]. The *seeds* example has appeared in [62] as a classic WINBUGS example and the dataset is originally from [21].

The model specification is as follows:

$$\beta_{2i} \sim \text{N}(0, \sigma_{RE}^2), \tag{2.19}$$

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_i + \beta_{2i}, \tag{2.20}$$

$$r_i \sim \text{Binom}(n_i, p_i), \tag{2.21}$$

where r_i is the binary outcome of interest for individual i, the x_i are the explanatory data collected per individual i, the n_i are the number of replications per individual i, and the β_{2i} are the unobserved random effects for individual i. We treat $x_1, ..., x_N$ as the observed random variables, $n_1, ..., n_N$ as fixed constants and the β_{2i} and p_i as latent variables. The parameters of interest are β_0, β_1, and σ_{RE}. Our dataset consists of twenty one individuals ($N = 21$).

We experiment with two initial values, $(\beta_0, \beta_1, \sigma_{RE}) = (0, 0, 1)$ and $(\beta_0, \beta_1, \sigma_{RE}) = (-1, -1, 4)$, as well as two different MCMC sample sizes, 20 and 300. With MCMC sample

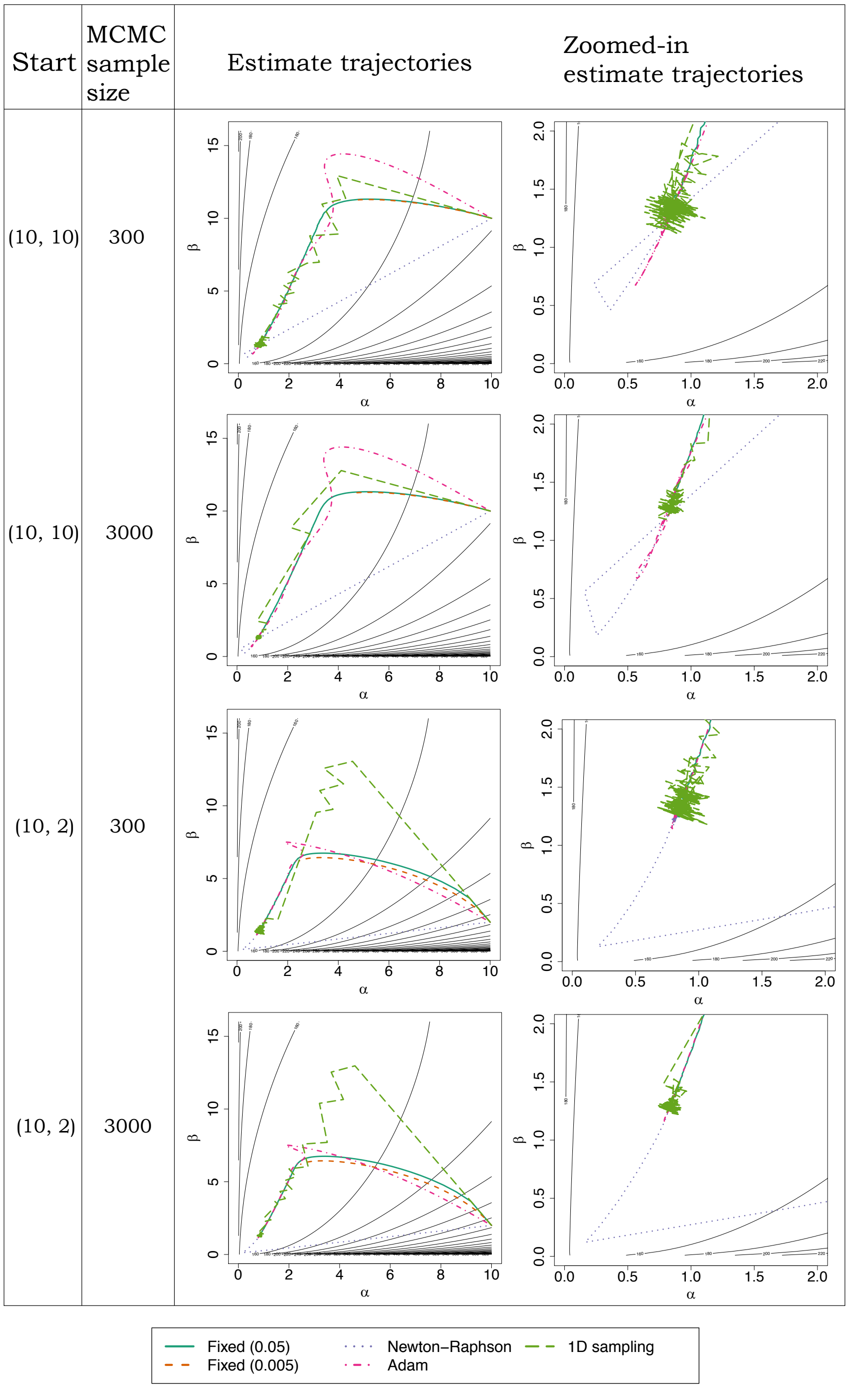


Figure 2.2: Estimate trajectories for the *pump* example. Zoomed-in trajectories are based on the final 100 iterations of the algorithms. The true MLE is $(\hat{\alpha}, \hat{\beta}) = (0.823, 1.262)$

size 300 and initial value $(-1,-1,4)$, Newton-Raphson seems to be stuck at the boundary constraints and fails to get close to the benchmark MLE, showing that Newton-Raphson could be sensitive to the initial value. Figure 2.3 display the trajectories of the iterates for each algorithm with MCMC sample size 300 and initial value $(0,0,1)$. The larger fixed step-size is sensitive to large gradients, leading to erratic jumps. Contrasting with the *pump* example, here the smaller step-size is preferable, suggesting that the fixed step-size method might require a careful step-size choice to achieve well-behaved trajectories. All of the other methods appear to converge within a narrow band around the true parameters fairly quickly. Adam and Newton-Raphson still perform well with MCMC sample sizes of 20 (Table A.9), but both the small fixed step-size and the large fixed step-size approaches have trouble with a small MCMC sample size (Table 2.3). Comparing Table A.13 with Table A.9, we observe that for smaller MCMC sample sizes more iterations are typically needed to reach convergence. Remarkably, Adam and 1D sampling arrive at a solution in seconds rather than minutes, much faster in terms of computational time than MCEM (Tables A.9 and A.13). The smaller fixed step-size approach also does fairly well. In this case we can also compare to the results given by the specialized method in the lme4 package [4] in R, and we see close agreement in the estimates.

2.5.3 Case study: salamander (crossed random effects model)

Our next example, *salamander* [47], features a GLMM with crossed random effects, which lead to challenging estimation of the random effects' variances. Let y_i be the observed outcome of whether salamander pair i successfully mated or not. For pair i, we use $F(i)$ and $M(i)$ to denote the corresponding female and male. Let $\text{REF}_{F(i)}$ and $\text{REM}_{M(i)}$ denote the random effects for female $F(i)$ and male $M(i)$. The model specification is as follows:

$$\text{REF}_{F(i)} \sim N(0,\sigma_F^2), \tag{2.22}$$

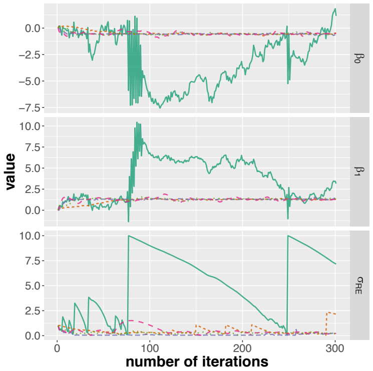
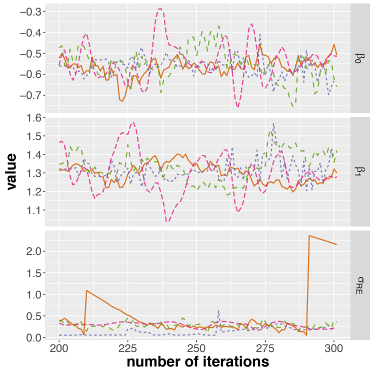
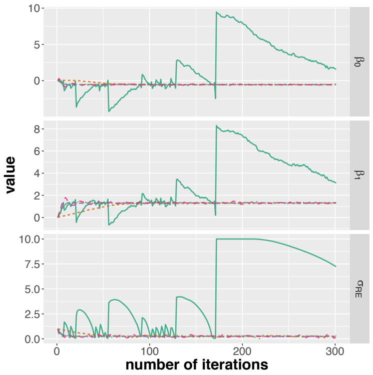
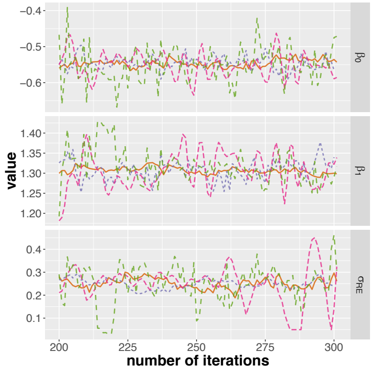
$$\text{REM}_{M(i)} \sim N(0,\sigma_M^2), \tag{2.23}$$

$$\text{logit}(\theta_i) = \beta_1\text{isRR}_i + \beta_2\text{isRW}_i + \beta_3\text{isWR}_i + \beta_4\text{isWW}_i + \text{REF}_{F(i)} + \text{REM}_{M(i)} \tag{2.24}$$

$$y_i \sim \text{Bern}(\theta_i), \tag{2.25}$$

where isRR, isRW, isWR, and isWW encode which population the female (first letter) and male (second letter) are from in each pair with R denoting “rough-butt” and W denoting “whiteside”; θ_i is the probability of mating for each pair i. We consider the random effects $\text{REF}_{F(i)}$, $\text{REM}_{M(i)}$ as latent variables. The top level parameters of interest are $\beta_1,\beta_2,\beta_3,\beta_4,\sigma_F^2$, and σ_M^2. Data from 360 pairs of salamanders ($N = 360$) is available in the *glmm* package [52] in R.

We experiment with two initial values of the six parameters $(\beta_1,\beta_2,\beta_3,\beta_4,\sigma_F^2,\sigma_M^2)$, $(2,2,2,2,2,2)$ and $(4,4,4,4,4,4)$. Our benchmark estimate is from the *lme4* package [4] in R. In addition, we also compare our estimates with the ones from the *glmm* package [52]. We take the advice in the documentation of *glmm* to increase the Monte Carlo sample size to 10^5 from the default of 10^4 in order to get a more reliable estimate of the parameters.

Start	MCMC sample size	Estimate trajectories	Zoomed-in estimate trajectories
(0, 0, 1)	20		
(0, 0, 1)	300		

— Fixed (0.05) - - Fixed (0.005) - - Newton–Raphson — Adam · · 1D sampling

Figure 2.3: Estimate trajectories for the *seeds* example. Zoomed-in trajectories are based on the final 100 iterations of the algorithms. The fixed step-size (0.05) trajectories are dropped in the zoomed-in plots for better resolution of the other methods. The *lme4* estimates suggest that a decent ML estimate should be around $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_{RE}) = (-0.548, 1.310, 0.249)$.

From Figure 2.4, we observe that all the methods arrive at a stable estimate of the β values quickly when the initial value is $(2, 2, 2, 2, 2, 2)$. In particular, Newton-Raphson and 1D sampling get close to the benchmark MLE within ten iterations. We compute the approximate log-likelihood values at the various MLEs via *lme4*. As shown in Table A.17 and A.19 in Supplementary Materials, the sampling based approaches yield MLEs close to the ones given by *glmm* and *lme4* in terms of both log-likelihood differences and MSE, regardless of the initial values.

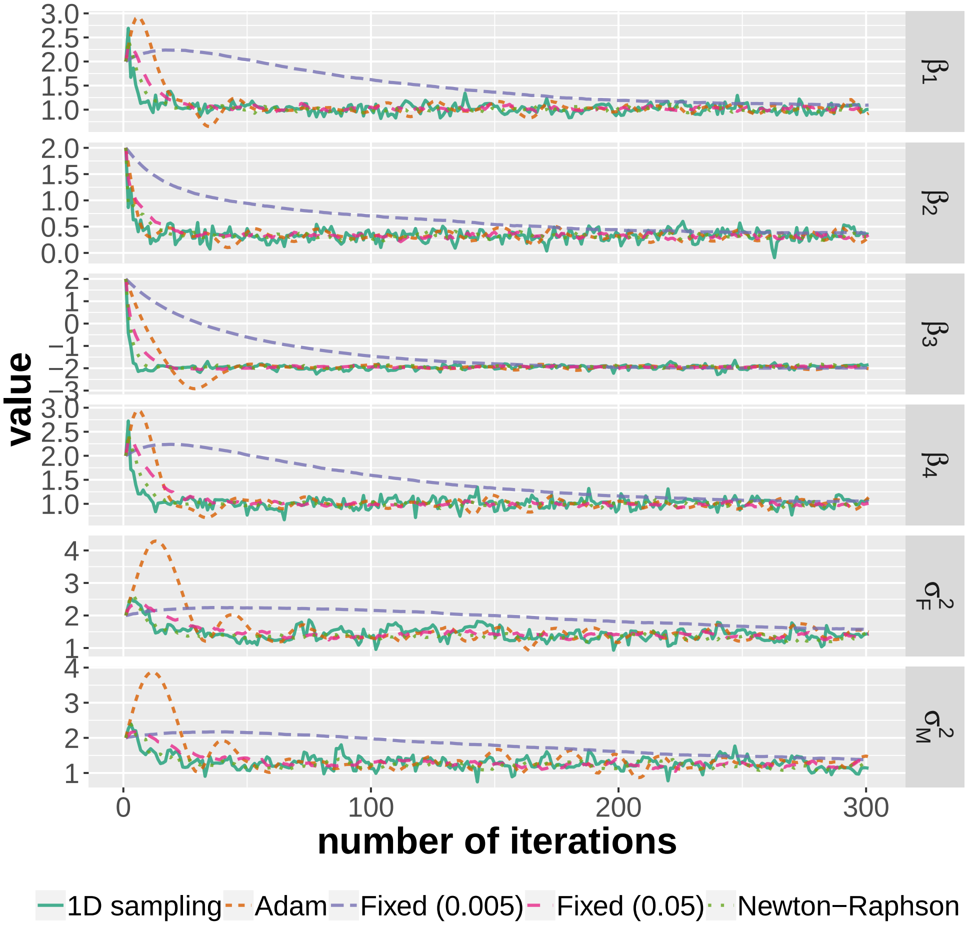


Figure 2.4: Estimate trajectories for the *salamander* example with MCMC sample size 300. Newton-Raphson and 1D sampling get close to the benchmark MLE within ten iterations, despite the fact that they do not pass the ad-hoc convergence criterion. The *glmm* estimates suggest that a decent ML estimate should be around $(\beta_1, \beta_2, \beta_3, \beta_4, \sigma_F^2, \sigma_M^2) = (1.023, 0.335, -1.908, 1.006, 1.326, 1.221)$, while the *lme4* estimates suggest that a decent ML estimate should be around $(\beta_1, \beta_2, \beta_3, \beta_4, \sigma_F^2, \sigma_M^2) = (1.008, 0.306, -1.896, 0.990, 1.174, 1.041)$.

2.6 Discussion

It is difficult, if not impossible, to establish theoretical comparisons among different sampling-based MLE approaches in terms of computational time. For example, while second-order method tends to converge in fewer iterations due to quadratic convergence, each computation of the Hessian matrix requires a considerable amount of time, so the benefit of converging in fewer iterations is being offset by the expensive Hessian computations. Due to difficulties in theoretical comparisons, experiments were conducted to investigate the performance of algorithms. Our experiments have two main limitations. The first limitation is that the conclusions might not be generalizable to all latent variable models. We have chosen examples where a benchmark MLE can be computed via specialized methods. For the *pump* example, we can explicitly find the marginal likelihood and compute the MLE; for the two GLMM example we use the results from *lme4* as benchmarks, although *lme4* relies on Laplace approximation and hence the results might not be accurate. The second limitation is that it is impossible to conduct experiments with all possible tuning parameter configurations for each of the MLE algorithms.

Empirically, the stochastic gradient approach is shown to be robust to small MCMC sample sizes. We conjecture that this robustness is due to the fast mixing behavior of the MCMC when we are sampling the latent variables given the data and the top-level parameters.

In our preliminary experiments we had experimented with other adaptive step-size schedules, Adagrad [25] and Adadelta [106]. We found that Adagrad decays the step-size too aggressively and results in slow convergence; Adadelta often ends in highly oscillating behaviors near convergence. Hence we do not include Adagrad and Adadelta in our work.

We observe in numerical experiments that HMSGD with Adam and 1D sampling typically work well with default tuning parameters. In addition, their estimate trajectories are relatively stable compared to HMSGD with fixed step-size. We summarize the challenges of existing MLE methods and their solutions via HMSGD-based methods in Table 2.1.

2.7 Conclusion

We presented a unifying framework for various sampling-based MLE algorithms and proposed various extensions. In particular, we have experimented with the use of adaptive step-size schedule Adam in the context of MLE for hierarchical models, and introduced the 1D sampling approach as a viable step-size determination procedure. Our numerical experiments have shown promising results for various algorithms, especially Adam and 1D sampling, in terms of achieving short computational time and obtaining reasonable parameter estimates. We have also found that automatic differentiation not only speeds up gradient computations but also helps stabilize Newton-Raphson method. We provide access to these algorithms in an easy-to-use format that is still customizable, and we hope these techniques can be valuable to practitioners in many fields.

Method	Challenge(s)	HMSGD-based solution
EM	Analytical derivation of E-step and/or M-step is required.	HMSGD-based solutions do not require analytical derivation.
MCEM	Fully-automated MCEM is computationally slow due to increasingly large MCMC sizes.	HMSGD with Adam can work with reasonably well with small MCMC sample sizes.
HMSGD with fixed step-size	It is sensitive to tuning and often leads to erratic estimate trajectories.	HMSGD with Adam/1D sampling typically works well with default tuning parameters and estimate trajectories are less susceptible to erratic jumps.
MCNR	Approximate Hessian is required and can be ill-conditioned. It might be sensitive to initial values.	HMSGD with Adam/1D sampling does not require Hessian and is less sensitive to initial values.

Table 2.1: Current challenges for existing MLE methods and their solutions via HMSGD-based approaches.

Acknowledgement

I would like to thank Chris Paciorek for his numerous suggestions in the project, in particular the idea of warm start in the MCMC sampling. I would also like to thank Nick Michaud for NIMBLE's implementation of MCEM.

Chapter 3

To rarefy or not to rarefy: statistical trade-offs of rarefying microbiome data

This chapter is based on joint work with Ulas Karaoz, William Fithian, and Perry de Valpine.

3.1 Introduction

The rapid development of high-throughput sequencing technologies facilitates the production of many valuable microbiome datasets, allowing insightful investigation of microbial communities. To make proper use of these datasets, adjusting for the varying library sizes (also known as *sequencing depths* or *read depths*) is crucial. Failure to do so might result in drawing the wrong scientific conclusions from the subsequent inferential procedures. One popular approach to data normalization is rarefaction (often referred as *rarefying*). The typical rarefaction procedure is as follows:

1. Specify a desired library size L^*.

2. Discard all the samples with library size L_i less than L^*.

3. Subsample all the samples with library size L_i greater than L^* to L^*. This sampling is done via sampling without replacement.

The desired library size L^* is often chosen to be the smallest observed library size $L_{\min}$ among the samples. In this case no samples will be discarded. However, often the smallest observed library size $L_{\min}$ is much smaller than the library sizes for most of the samples, possibly due to sampling or processing failures. The data would become too noisy if all the samples were to be rarefied to $L_{\min}$. In this case researchers pre-specify a certain threshold L^* and proceed with the rarefaction procedure.

According to Willis [101], rarefaction were first proposed in Sanders [83] for alpha diversity comparisons in oceanography. As described in Weiss et al. [99], recently the practice of rarefaction has been applied in the context of studying beta diversity [41, 46]. From McMurdie and Holmes [64], “rarefying was first recommended for microbiome counts in order to moderate the sensitivity of the UniFrac distance [60] to library size, especially differences in the presence of rare OTUs (operational taxonomic units) [61]”.

There has been ongoing debate on whether rarefaction is a statistically justifiable procedure. Rarefying microbiome data has been criticized a statistically inadmissible procedure under the assumption that the data can be modeled with a negative binomial distribution [64]. Roughly speaking, the phrase “statistically inadmissible” means that there exists a statistical decision rule that has a lower statistical risk. Since rarefaction discards valuable data and leads to an inadmissible procedure, McMurdie and Holmes [64] recommend that rarefaction should never be used in practice. However, the mathematical argument on inadmissibility in [64] relies on the model assumption being correct, and McMurdie and Holmes [64] do not consider randomization tests for differential abundance testing in simulation studies. Willis [101] explains how rarefaction can introduce bias in the context of studying alpha diversity. On the other side of the debate, Weiss et al. [99] argue that “rarefying is still a useful normalization technique” since from simulation studies rarefying seems to be a more effective procedure to reduce the effect of uneven library sizes in inferential procedures.

To the best of our knowledge, previous studies in understanding rarefaction for statistical inference has been limited to simulation studies and/or specific data generating processes and modeling assumptions. We provide motivating examples in which without rarefaction statistical inference can be compromised. Our work develops a formal statistical framework for understanding rarefaction. To formally understand how rarefaction helps normalize the data, we propose a nonparametric graphical model for grouped microbiome data and clarify how various sources of randomness can affect inferential procedures for microbial compositions. In particular, the variance of the sample relative abundance can be decomposed into latent variation in microbial compositions and measurement error. Using a nonparametric graphical model framework not only allows us to make general claims about rarefying but also highlights the minimal assumptions for rarefying to be a justifiable procedure.

We establish that rarefaction guarantees the validity of permutation tests under the nonparametric graphical model, opening up the use of flexible testing procedures via permutations. In particular, rarefaction preserves the conditional multinomial distribution of the count data the while eliminating the potential dependence between library sizes and group memberships. While model-based approaches are gaining popularity as they often lead powerful procedures, inferences drawn from these approaches are often questionable since the underlying parametric assumptions are tricky to verify. If observations are exchangeable under the null hypothesis that there are no between group differences (referred as *null exchangeability*), one can use a model-based test statistic with permutation inference to guard against potential violations of model assumptions.

We develop rarefaction efficiency index (REI) as an indicator for efficiency loss, motivated by our theoretical considerations under the nonparametric graphical framework. REI can be viewed as an estimate of statistical efficiency between testing based on original data and testing based on rarefied data. We illustrate the use of REIs with a microbiome data example, and simulation studies show that REIs are indeed informative about the sensitivity loss from rarefaction. We also provide simulated examples in which, without rarefaction, common statistical procedures for microbiome data (PERMANOVA and DESeq2) can have inflated Type I error rate.

Throughout the rest of the paper, we assume we have grouped microbiome data. To simplify the discussion, we assume there are only two groups, although our discussion can be easily extended to the case of multiple groups.

3.2 Notations

Consider a grouped microbiome dataset with J OTUs of interest and samples can be categorized into $G = 2$ different groups. For group $g \in \{1, 2\}$, there are n_g observations. For $i \in \{1, ..., n_g\}$, let $\mathbf{x}_i^{(g)} \in \mathbb{N}^J$ denote the ith count vectors in group g so that the kth entry of the vector corresponds to the raw count of the kth OTU. Our data consist of two groups of observations $\{\mathbf{x}_1^{(1)}, ..., \mathbf{x}_{n_1}^{(1)}\}$ and $\{\mathbf{x}_1^{(2)}, ..., \mathbf{x}_{n_2}^{(2)}\}$. The library size for sample i in group g is $L_i^{(g)} = \sum_{j=1}^J x_{ij}^{(g)}$. Let L^* be the rarefied depth, and $\{\mathbf{x}_1^{(1)*}, ..., \mathbf{x}_{n_1}^{(1)*}\}$, $\{\mathbf{x}_1^{(2)*}, ..., \mathbf{x}_{n_2}^{(2)*}\}$ be the rarefied data. We assume each replicate has an unobserved latent composition $\boldsymbol{\pi}_i^{(g)}$, which is a random vector of OTU relative abundances. For positive integer K, we use $\mathbf{1}_K$ is a vector of K ones.

When we narrow our discussion for n samples in the same group g for a particular OTU, for $i \in \{1, ..., n\}$, let x_i be the raw count of the particular OTU in sample i, L_i be the associated library sizes, and π_i be the latent relative abundance. We assume all the samples share a common expected OTU relative abundance p. We further drop the subscript i when discussing a generic observation.

3.3 Motivating examples

We carry out simulations to illustrate how varying library size distributions, as well as various violations of model assumptions, can potentially undermine popular statistical procedures in microbiome studies. In particular, we study the Type I error control for (1) the permutation t-test of species richness, and (2) DESeq2, a differential abundance testing approach.

3.3.1 Permutation t-test for species richness

To illustrate how different library sizes can compromise statistical inference, we provide a simple example of comparing species richness between two groups of samples with the same underlying microbial composition distribution. Intuitively, observations with larger library sizes have larger species richness solely because more microbes are being sampled. We

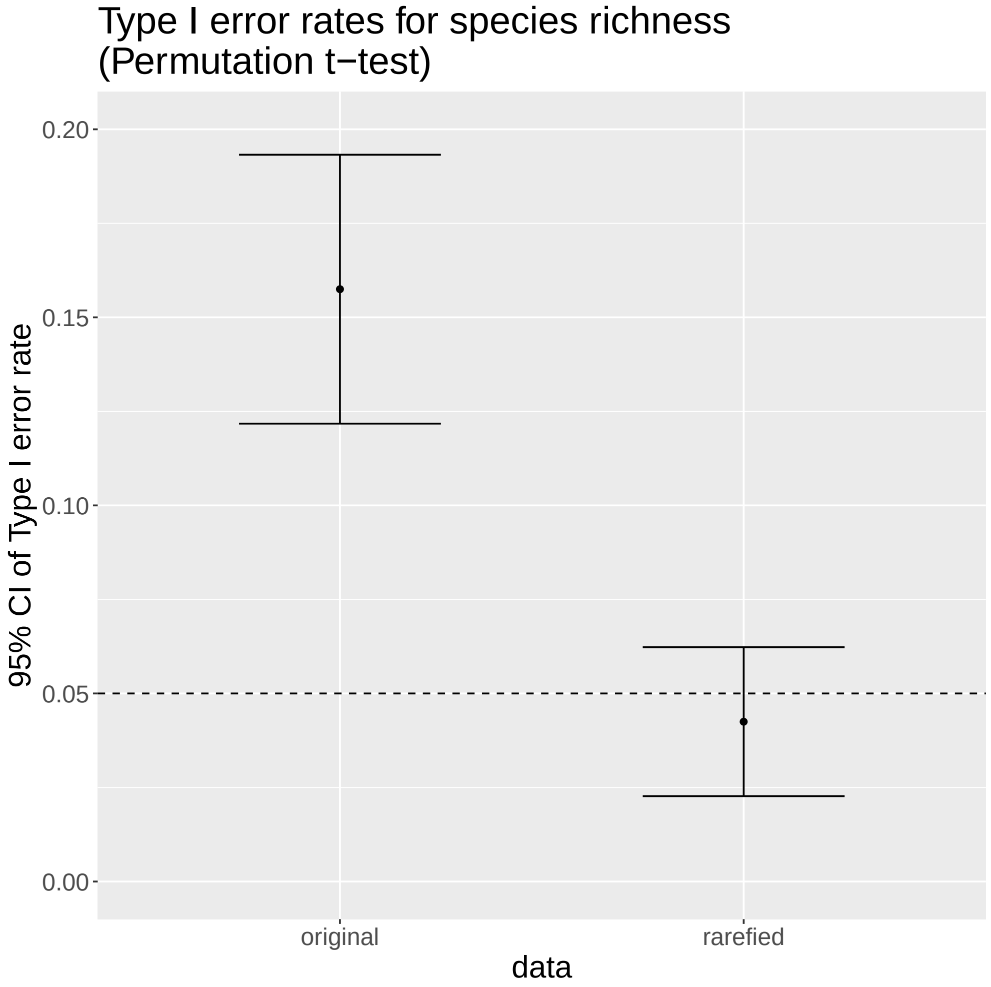


Figure 3.1: 95% confidence intervals of the Type I error rates for the permutation t-test on species richness. The permutation t-test rejects the null hypothesis that the two groups of observations have the same mean species richness if the permutation p-value is less than 5%. Simulations show that testing based on rarefied data properly control the Type I error rate.

simulate 200 datasets from the following: for group $g = 1, 2$ and observation $i = 1, ..., 30$,

$$\boldsymbol{\pi}_i^{(g)} \sim \text{Dirichlet}(\boldsymbol{\alpha}), \tag{3.1}$$

$$L_i^{(g)} \sim P_g, \tag{3.2}$$

$$\mathbf{x}_i^{(g)} | L_i^{(g)} \sim \text{Multinomial}(L_i^{(g)}, \boldsymbol{\pi}_i^{(g)}), \tag{3.3}$$

where $\boldsymbol{\alpha}$ is the empirical vector of relative abundances (the number of OTUs $J = 9719$) from the microbiome preservation dataset [85], scaled by a concentration parameter estimated from the data; P_1 and P_2 are empirical distributions of library sizes for observations subject to the freezethaw treatment and observations subject to the 4°C storage treatment respectively. Note that the distribution of the underlying microbial compositions for two groups are the same, so a test that rejects the null hypothesis the two groups have the same species richness commits a Type I error.

Figure 3.1 shows that two 95% confidence intervals for the Type I error rates of the permutation t-test (number of permutations = 200; the test rejects the null if the permutation p-value fall below 5%) of species richness, one for original data and one for rarefied data (that is, data are rarefied to the smallest library size). Based on the confidence intervals, the Type I error is being inflated if original data are being used, and it is properly controlled if rarefied data are being used.

3.3.2 DESeq2

DESeq2 [59] is a negative binomial-based method to test for differential gene expression. Since its introduction, it has garnered popularity in microbiome studies in differential abundance testing [35, 45, 94, 103, 105].

We first investigate the robustness of DESeq2 against different data conditions. For each of the following conditions, we generate 100 datasets with $J = 2000$ OTU:

1. baseline: $n_1 = n_2 = 30$, and for $g \in \{1, 2\}$,

$$\boldsymbol{\pi}_i^{(g)} \sim \text{Dirichlet}\left(30\frac{\mathbf{1}_J}{J}\right), \tag{3.4}$$

$$L_i^{(g)} \sim \text{Poisson}(10000), \tag{3.5}$$

$$\mathbf{x}_i^{(g)} | L_i^{(g)} \sim \text{Multinomial}(L_i^{(g)}, \boldsymbol{\pi}_i^{(g)}); \tag{3.6}$$

2. small sample sizes: same as the baseline model except $n_1 = n_2 = 10$;

3. different mean library sizes: same as the baseline model except $L_i^{(2)} \sim \text{Poisson}(50000)$;

4. mixture: same as the baseline model except $\boldsymbol{\pi}_i^{(g)}$ comes from a mixture of two Dirichlet distributions:

$$\boldsymbol{\pi}_i^{(g)} \sim 0.3\text{Dirichlet}\left(30\frac{\mathbf{1}_J}{J}\right) + 0.7\text{Dirichlet}\left(3[1.5\frac{\mathbf{1}_{J/2}}{J}, 0.5\frac{\mathbf{1}_{J/2}}{J}]\right).$$

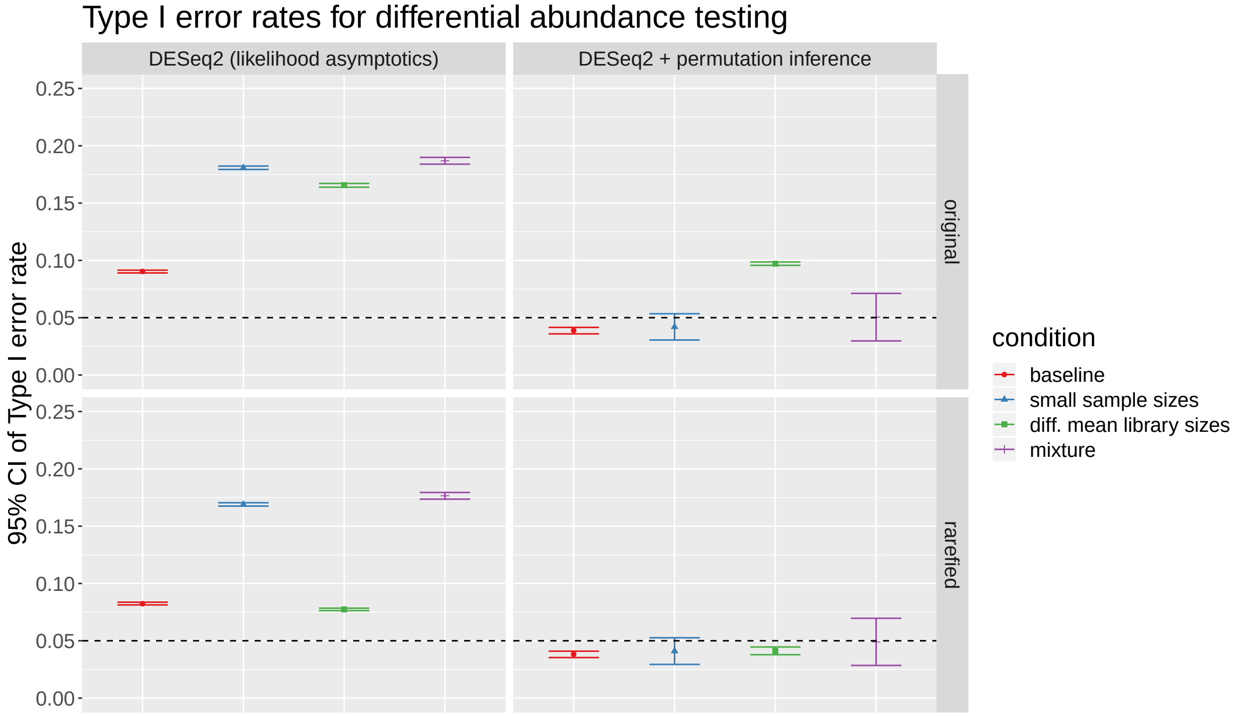


Figure 3.2: 95% confidence intervals of DESeq2 Type I error rates in differential abundance testing based on 2000 OTUs and 100 simulated datasets under four difference data generating conditions. Top left: DESEq2 on original data; top right: permutation test with DESEq2 test statistic on original data; bottom left: DESeq2 on rarefied data; bottom right: permutation test with DESEq2 test statistic on rarefied data. In each condition, the null hypothesis that there are no compositional difference holds, so the true Type I error rate for each OTU is 5%. Simulations show that permutation tests on rarefied data control the Type I error rate approximately at the correct 5% level.

Ww have two versions of each dataset, original and rarefied. For each version of the dataset, we apply DESeq2 with two modes of testing: a) likelihood asymptotics: use the default Wald test to obtain p-values for each OTU; b) permutation: p-values are computed based on permutations for each OTU. For each OTU, we reject the null hypothesis that the OTU is not differentially abundant if the p-value falls below 0.05. Note that in all the simulations the null hypothesis holds.

Figure 3.2 shows the 95% confidence intervals of Type I error rates in each scenario. Inferences based on asymptotics tend to have an inflated Type I error rate in all simulations. On the other hand, inferences based on permutations tend to achieve the expected Type I error rate except when the mean library sizes for the two groups differ. Regardless of the conditions, the combination of rarefying and permutations consistently yields approximately the expected 5% Type I error rate. In theory, we expect the confidence interval should cover the expected 5% Type I error rate; however, very small deviations arises due to DESeq2 not being able to estimate OTU count dispersions for certain permuted samples.

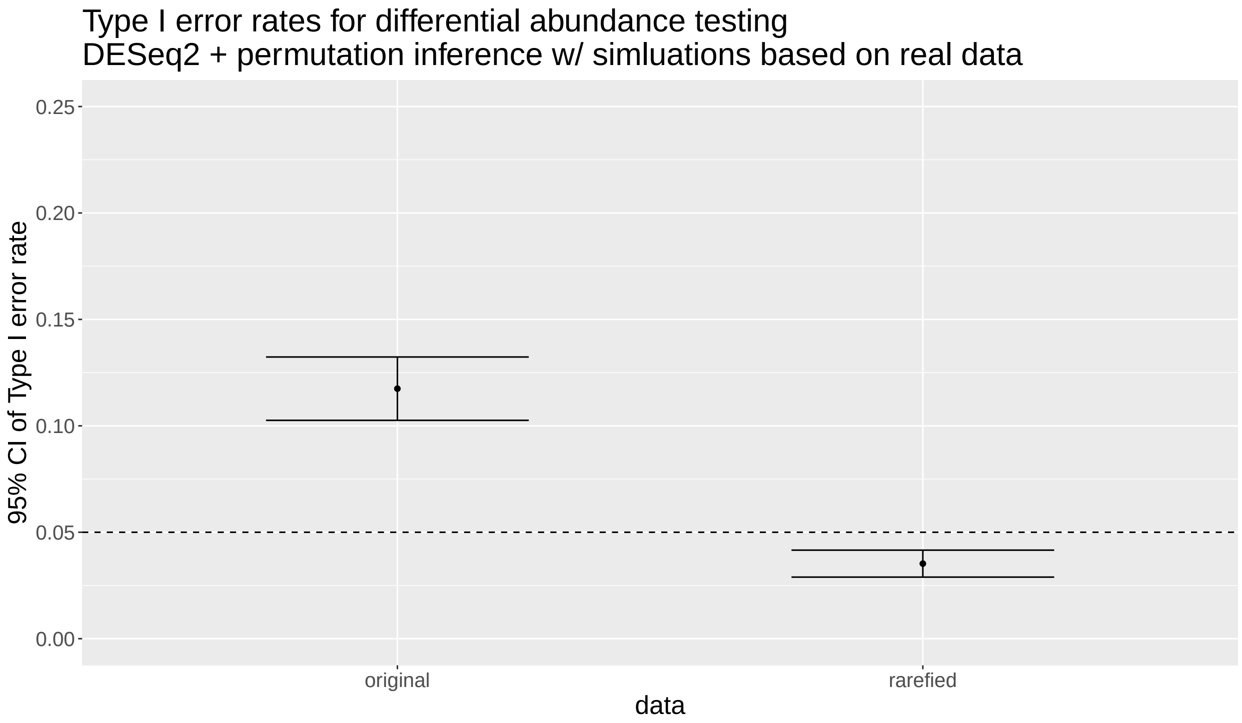


Figure 3.3: 95% confidence intervals of DESeq2 Type I error rates in differential abundance testing based on 100 simulated datasets as close to the microbiome preservation dataset as possible. Since the latent compositions are simulated from the same distribution, the true Type I error rate for each OTU is 5%. Simulations show that permutation tests on rarefied data control the Type I error rate approximately at the correct 5% level.

We also study the robustness of DESeq2 under a setup in which simulated data are as close to an actual dataset as possible. We choose the actual dataset to be the micriobiome preservation dataset [85]. Let P_1 and P_2 are empirical distributions of library sizes for observations subject to the freezethaw treatment and observations subject to the 4°C storage treatment respectively. Let Q be the empirical distributions of relative abundance vectors for observations subject to the freezethaw treatment. Each of the simulated datasets is generated as follows:

1. For group $g = 1$, simulate 30 library sizes $L_1^{(1)}, ...L_{30}^{(1)}$ from P_1.
2. For group $g = 2$, simulate 30 library sizes $L_1^{(2)}, ...L_{30}^{(2)}$ from P_2.
3. For $g = 1, 2$ and $i = 1, ..., 30$, simulate the latent composition $\pi_i^{(g)}$ from Q and $x_i^{(g)}$ from Multinomial$(L_i^{(g)}, \pi_i^{(g)})$.

Figure 3.3 shows the 95% confidence intervals of Type I error rates under this simulation setup for original data and rarefied data, using DESeq2 with permutational inference. Testing based on original data yield inflated Type I error rate, while testing based on rarefied data does not.

3.4 Framework

To facilitate a rigorous discussion of rarefaction, we develop a nonparametric statistical framework for microbiome data. The setup is that there are multiple subjects from each group, and each subject has an associated composition of the microbial community, say $\boldsymbol{\pi}_i$ for the i-th subject. Due to limitations of sequencing procedures, the specimen extracted from a subject comes from a sampling of the microbial community, and typically the total count (library size) of the specimen is not informative about the absolute total abundance of the specimen.

To provide intuition of this setup, suppose we collect two groups of soil samples. Each group of soil samples comes from a different region. Within each group, the samples can have different microbial compositions, perhaps due to unobserved conditions of each sample such as temperature and humidity at sample collection, or perhaps due to inherent variation across identical conditions. The microbial composition of a sample is never entirely observed; for each sample, we take a swab and perform the sequencing procedure. This can be viewed as subsampling the microbial community in the soil sample.

Motivated by the setup, we propose the following nonparametric model for microbiome data:

$$\text{latent composition:} \qquad \boldsymbol{\pi}_i^{(g)}|g \sim f_{\boldsymbol{\pi}}(\cdot|g) \tag{3.7}$$

$$\text{library size:} \qquad L_i^{(g)}|\boldsymbol{\pi}_i^{(g)}, g \sim f_L(\cdot|\boldsymbol{\pi}_i^{(g)}, g) \tag{3.8}$$

$$\text{count:} \qquad \mathbf{x}_i^{(g)}|\boldsymbol{\pi}_i^{(g)}, L_i^{(g)} \sim \text{Multinomial}(L_i^{(g)}, \boldsymbol{\pi}_i^{(g)}), \tag{3.9}$$

where $f_{\boldsymbol{\pi}}(\cdot|g)$ is a probability density function supported on the p-dimensional probability simplex and $f_L(\cdot|\boldsymbol{\pi}_i^{(g)}, g)$ is a probability mass function supported on non-negative integers. For a given group, there is an underlying mean composition, and values of $\boldsymbol{\pi}_i^{(g)}$ represent latent variation in sample composition. The multinomial distribution is a natural modeling assumption if the sampling of microbial community, from data collection to obtaining the raw counts, is assumed to be representative of the underlying microbial composition.

In the graphical model the arrows do not necessarily imply causal relationships. A generic joint distribution of $(\boldsymbol{\pi}_i^{(g)}, L_i^{(g)})|g$ can be described by $f_{\boldsymbol{\pi}}(\cdot|g)$ and $f_L(\cdot|\boldsymbol{\pi}_i^{(g)}, g)$ since the following identity holds

$$f_{\boldsymbol{\pi},L}(\boldsymbol{\pi}_i^{(g)}, L_i^{(g)}|g) = f_{\boldsymbol{\pi}}(\boldsymbol{\pi}_i^{(g)}|g) f_L(L_i^{(g)}|\boldsymbol{\pi}_i^{(g)}, g) \tag{3.10}$$

for any joint distribution $f_{\boldsymbol{\pi},L}(\cdot,\cdot|g)$.

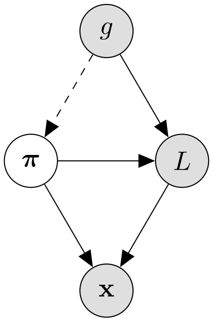


Figure 3.4: A nonparametric graphical model for grouped microbiome data. In the model, g is the group the observation belongs to, $\boldsymbol{\pi}$ is the (latent) microbial composition, L is the library size, and $\mathbf{x}$ is the vector of OTU counts. The shaded nodes are observed quantities. The dashed arrow represents the hypothesis of interest: whether microbial compositions vary across different groups.

3.4.1 Example: Dirichlet-Multinomial

Dirichlet-Multinomial (DM) models have been used in microbiome studies to model the count vector of multiple taxa [55, 39, 15, 17, 93]. In the DM model, the latent composition follows a Dirichlet distribution and the conditional distribution of the count vector follows a multinomial distribution. Dirichlet distribution is a conjugate prior for the multinomial distribution, which greatly simplifies subsequent computations.

For testing differential abundance in a single OTU, the negative binomial (NB) distribution is frequently used to model counts. When the count vectors of microbial communities are modeled by the DM model, the counts for each OTU follow approximately the NB distribution. See Supplementary Materials for more details.

3.5 How much does rarefaction hurt statistical inference?

Rarefaction has been criticized for wasting data since we effectively remove a portion of the data in the downsampling procedure [64]. The subsampling in rarefying will inevitably increase the variance of test statistics and decrease the power of subsequent testing procedures. The key question is how much rarefaction hurts statistical inference. We address this question through theoretical considerations and simulation studies. We focus our discussion on two relatively simple testing procedures, the negative binomial Wald test and the two-sample t-test. More involved statistical procedures such as DESeq2 [59] and permutational analysis of variance (PERMANOVA, [2]) are hard to analyze theoretically because they integrate various statistical principles. Negative binomial-based tests are used in two popular microbiome analysis routines, edgeR [79] and DESeq2 [59], while the two-sample t-test (and the closely related z-test) is a classical statistical test widely used in microbiome studies [100, 104], and it forms an integral part of more involved approaches in the popular package *limma* [77, 71] for differential expression analysis.

3.5.1 Theoretical considerations

We summarize the key theoretical findings in this section and leave all the mathematical details and the precise statements in Supplementary Materials.

Under the proposed graphical model framework, it can be shown that rarefaction to a pre-specified depth leads to the exchangeability of observations under the null hypothesis, which is required for permutation tests to achieve correct Type I error rates. The intuition is that the effects of group memberships and latent composition on library sizes are removed by reducing all the library sizes to the same number.

It can further be shown that the variance of sample relative abundance within each group can be written as the sum of latent variation and measurement error:

$$\operatorname{Var}\left(\frac{x}{L}\right) = \underbrace{\operatorname{Var}(\pi)}_{\text{latent variation}} + \underbrace{\mathbb{E}\left[\frac{\pi(1-\pi)}{L}\right]}_{\text{measurement error}}. \tag{3.11}$$

The first term in $\operatorname{Var}\left(\frac{x}{L}\right)$ is the variance of the latent relative abundance π, representing the *latent variation.* The latent variation arises from having multiple observations belonging to the same group. Samples from the same group do not necessarily have the same OTU proportion due to individual differences. The latent variation $\operatorname{Var}(\pi)$ is determined by $f_{\boldsymbol{\pi}}(\cdot|g)$ in the graphical model. The second term in $\operatorname{Var}\left(\frac{x}{L}\right)$ is the *measurement error*, arise from the sequencing procedure of each sample. Intuitively if the library size is larger, the precision increases and hence the measurement error decreases.

Statistical efficiency is often used to compare hypothesis testing procedures. Roughly speaking, a more efficient testing procedure requires fewer observations to achieve the same power. It can be shown that the (asymptotic) statistical efficiency (measured by asymptotic relative efficiency, ARE) based on original data versus rarefied data for the two-sample t-test is approximately

$$\frac{\frac{1}{n_1}\operatorname{Var}\left(\frac{x^{(1)}}{L^{(1)}}\right) + \frac{1}{n_2}\operatorname{Var}\left(\frac{x^{(2)}}{L^{(2)}}\right)}{\frac{1}{n_1}\operatorname{Var}\left(\frac{x^{(1)*}}{L^{*}}\right) + \frac{1}{n_2}\operatorname{Var}\left(\frac{x^{(2)*}}{L^{*}}\right)} \tag{3.12}$$

and the ARE for the negative binomial Wald test is upper bounded by the above quantity (3.12). See Appendix B.6 for details.

3.5.2 Rarefaction efficiency index

Consider the setup for two groups and J OTUs. It would be helpful to have an index that informs us whether rarefaction would lead to a significant loss in statistical efficiency. We call the quantity in (3.12)

$$\frac{\frac{1}{n_1}\mathrm{Var}\left(\frac{x^{(1)}}{L^{(1)}}\right)+\frac{1}{n_2}\mathrm{Var}\left(\frac{x^{(2)}}{L^{(2)}}\right)}{\frac{1}{n_1}\mathrm{Var}\left(\frac{x^{(1)*}}{L^{*}}\right)+\frac{1}{n_2}\mathrm{Var}\left(\frac{x^{(2)*}}{L^{*}}\right)}$$

the *rarefaction efficiency index* (REI).

In practice we would have to estimate the variance terms in the REI. To formalize ideas, let $x_{ij}^{(g)}$ and $L_i^{(g)}$ be the observed count for OTU j in sample i from group g and the associated library size respectively. For group g and OTU j, we can directly estimate $\mathrm{Var}(x_{ij}^{(g)}/L_{ij}^{(g)})$ using the sample variance of the observed relative abundances,

$$S_j^{(g)} := \frac{1}{n_g - 1}\sum_{i=1}^{n_g}\left(\frac{x_{ij}^{(g)}}{L_i^{(g)}} - \frac{1}{n_g}\sum_{l=1}^{n_g}\frac{x_{ij}^{(g)}}{L_l^{(g)}}\right)^2. \tag{3.13}$$

While we can estimate $\mathrm{Var}(x_{ij}^{(g)*}/L^*)$ using the sample variance of rarefied data, the realized value of this estimator can be larger than $S_j^{(g)}$. Thus we use an alternative estimator for $\mathrm{Var}(x_{ij}^{(g)*}/L^*)$: $S_j^{(g)} + V_j^{(g)}(L^*)$, with

$$V_j^{(g)}(L^*) := \frac{1}{n_g L^*}\sum_{i=1}^{n_g}\frac{x_{ij}^{(g)}}{L_i^{(g)}}\left(1 - \frac{x_{ij}^{(g)}}{L_i^{(g)}}\right)\left(\frac{L_i^{(g)} - L^*}{L_i^{(g)} - 1}\right), \tag{3.14}$$

which can be shown to be the estimated additional variance induced by rarefying data to depth L^* (see Appendix B.8). The *sample rarefaction efficiency index* (sample REI) of OTU j at rarefied depth L^* is defined as follows:

$$\widehat{\mathrm{REI}}_j(L^*) := \frac{\frac{1}{n_1}S_j^{(1)} + \frac{1}{n_2}S_j^{(2)}}{\frac{1}{n_1}(S_j^{(1)} + V_j^{(1)}(L^*)) + \frac{1}{n_2}(S_j^{(2)} + V_j^{(2)}(L^*))}. \tag{3.15}$$

Since $V_j^{(1)}(L^*)$ and $V_j^{(2)}(L^*)$ are always nonnegative, $\widehat{\mathrm{REI}}_j(L^*)$ is always less than or equal to 1. For simplicity, we define the sample REI for a dataset to be the average of sample REIs over all OTUs, but in practice one might opt for a weighted average if OTUs have varying importance in the analysis.

A sample REI close to 1 suggests that inference based on rarefied data is almost as efficient as inference based on original data. The idea is that when latent variation in composition is large relative to measurement error, the sample REI is close to 1 and there is not much loss of information from rarefying. On the other hand, a sample REI close to 0 suggests a huge loss in sensitivity when data are rarefied.

3.5.2.1 An usage example of rarefaction efficiency index

To illustrate how REIs can be applied in practice, we consider a dataset in a study on the effects of preservation and storage conditions on the fecal microbiomes [85]. The rarefied depth is chosen to be 30000. Suppose we select observations preserved for four weeks with 95% ethanol and storage temperature being either 20°C or 4°C. We group the observations by the storage temperature. The sample REI for comparing these two groups is found to be 0.76. On the other hand, suppose we select observations from two particular human subjects. We group the observations by the subjects. The sample REI for comparing these two groups is found to be 0.61.

Intuitively, if observations are grouped by treatments, we expect the latent variation to be large since observations come from different individuals; consequently, the loss in efficiency due to rarefaction should be small and the sample REI should be large. On the other hand, if observations are grouped by subjects, we expect the latent variation to be small since all the observations come from the same individual and are subject to preservation treatments to a certain extent; consequently, the loss in efficiency due to rarefaction should be large and the sample REI should be small.

Suppose researchers are willing to tolerate up to 25% loss in statistical efficiency. Based on the sample REIs, researchers should be concerned about using rarefaction for comparisons between the two subjects, but less concerned for comparisons between the two treatments.

3.5.3 Simulation studies

We carry out simulation studies to (1) investigate the impact of latent composition variation on the power of testing procedures and (2) see if REI properly reflects the discrepancy in power between original data and rarefied data. To echo the theoretical considerations, we focus on the two-sample t-test and the negative binomial based Wald test in single OTU differential abundance testing for two groups of samples. In the simulations, the concentration parameter α ranges over three values $\{100, 1000, 10000\}$. A smaller α represents more overdispersion; that is, more latent composition variation. The two groups have the same number of observations n, and the sample size is either $n = 20$ or $n = 100$. To keep the simulation as realistic as possible, the library size distributions are derived from an actual dataset: for group 1, the library size distribution (denoted by P_1) is the empirical library size distribution based on the Human Microbiome Project dataset (HMPv35, [42]); for group 2, the HMPv35 library sizes are multiplied by a factor of either 2 or 10 (denoted by P_2). We screen out all the rare species in the HMPv35 dataset (the number of remaining OTUs is $J = 366$). We use the empirical relative abundances, denoted as $\mathbf{v}_1$, as the true baseline relative abundances in the simulation. The empirical relative abundance of the first OTU is 0.1%.

We simulate 500 datasets from the following Dirichlet-Multinomial model: for the first group, the model is $\boldsymbol{\pi}_i^{(1)} \sim \text{Dirichlet}(\alpha \mathbf{v}_1)$, $L_i^{(1)} \sim P_1$, $\mathbf{x}_i^{(1)} | L_i^{(1)} \sim \text{Multinomial}(L_i^{(1)}, \boldsymbol{\pi}_i^{(1)})$; for the second group, we introduce a fold change factor (FC) for the first OTU, and the model is $\boldsymbol{\pi}_i^{(2)} \sim \text{Dirichlet}(\alpha \mathbf{v}_2), L_i^{(2)} \sim P_2$, $\mathbf{x}_i^{(2)} | L_i^{(2)} \sim \text{Multinomial}(L_i^{(2)}, \boldsymbol{\pi}_i^{(2)})$, where $\mathbf{v}_2 = [\mathbf{v}_{1,1}(FC), [1 - v_{11}(FC)]\mathbf{v}_{1,2:J}]$, with $\mathbf{v}_{1,1}$ the relative abundance of the first OTU, and $\mathbf{v}_{1,2:J}$ the relative abundances of the rest of the OTUs. For the rarefied depth, we simply take the smallest observed library size.

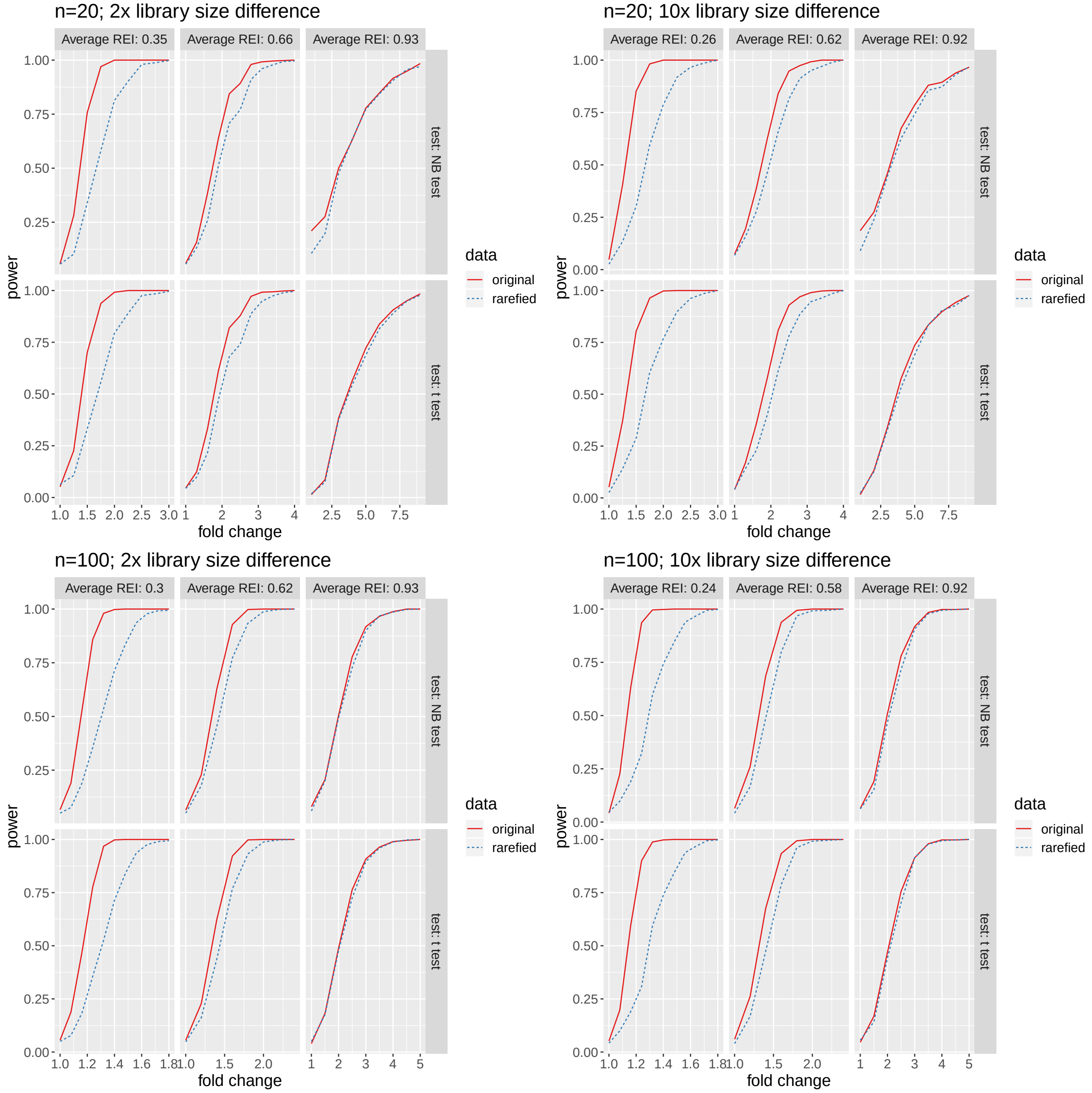


Figure 3.5: Power curves of differential abundance testing from simulation studies. Datasets are simulated from a Dirichlet-Mutlinomial model. In Group 1, the relative abundance of the OTU of interest is 0.1%. Within each figure, panels are arranged from left to right based on increasing overdispersion (i.e. decreasing α).

We test for differential abundance for the first OTU. The results for the NB test and the two-sample t-test are similar, although in general the power from NB test is slightly higher. Figure 3.5 shows that both tests control the Type I error at around the 5% level when the fold change is 1 (i.e.: there is no difference in the relative abundance of the first OTU for the two groups) except for the scenario in which the sample size is small and the overdispersion is large.

In the overdispersed case $\alpha = 100$, we can see from Figure 3.5 that the power curves for tests based on original data are very similar to those based on rarefied data, regardless of the sample sizes or the difference in mean library sizes between the two groups. As the concentration parameter α increases (in other words, the latent composition variation decreases), the gap between the original power curve and the rarefied one becomes more prominent in each of the scenarios.

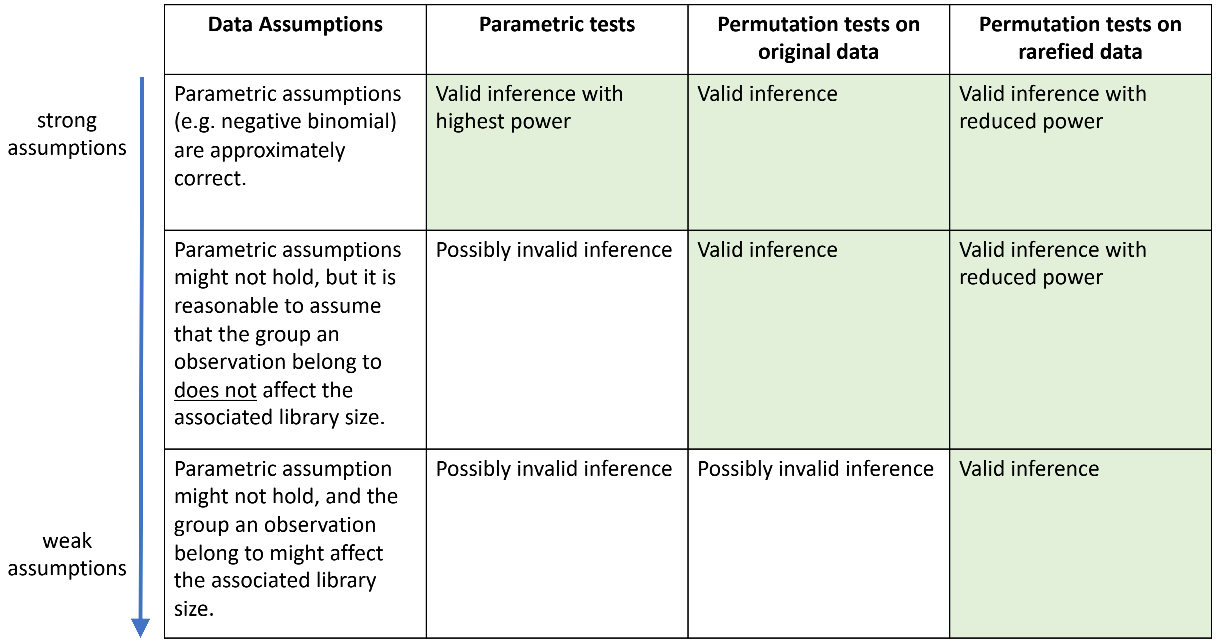
Comparing the plots in Figure 3.5, we can see that rarefying worsens the power more significantly as the mean library size differences increases. This is expected because the downsampling in rarefying becomes more aggressive for the larger library sizes. resulting in a bigger loss in precision.

From Figure 3.5, we observe that whenever the sample REI is small, the gap between the power curves is large, regardless of sample sizes. When the sample REI is around 0.9, the power curves essentially overlap, meaning that the effect of rarefaction on sensitivity is negligible; when the sample REI is around 0.7, the gap between power curves is visible but small. In practice, the default threshold for the sample REI can be set to 0.7; if the sample REI is below 0.7, one must beware of the drop in sensitivity due to rarefaction.

3.6 Discussion

We have shown that rarefaction with permutation tests provides robustness to differential abundance testing. Analysts often face the dilemma of whether to rarefy or not to rarefy. If the analyst is certain library sizes are independent of groups, rarefaction is not necessary since the observations are exhcangeable under the null hypothesis and permutation tests would yield valid p-values. To check if library sizes are independent of group memberships, one may use a parametric test such as ANOVA, or a nonparametric test such as Kruskal-Wallis test, to test for differences in library sizes among groups. In practice, analysts might want to err on the side of caution and opt for rarefaction at the potential cost of sensitivity loss; the extent of such sensitivity loss can be informed by REIs.

As long as observations are exchangeable under the null hypothesis, permutation tests are always valid regardless of the choice of the test statistic. However, permutation tests have several drawbacks. First, permutation tests are computationally expensive, especially when they are used in conjunction with a model-based approach. This problem is somewhat alleviated with parallel computing and the ever increasing computational power. Second, it can be challenging to deal with complex experimental designs in microbiome studies with permutation tests due to the exchangeability assumption underlying such tests.



	Data Assumptions	Parametric tests	Permutation tests on original data	Permutation tests on rarefied data
strong assumptions	Parametric assumptions (e.g. negative binomial) are approximately correct.	Valid inference with highest power	Valid inference	Valid inference with reduced power
	Parametric assumptions might not hold, but it is reasonable to assume that the group an observation belong to does not affect the associated library size.	Possibly invalid inference	Valid inference	Valid inference with reduced power
weak assumptions	Parametric assumption might not hold, and the group an observation belong to might affect the associated library size.	Possibly invalid inference	Possibly invalid inference	Valid inference

Figure 3.6: Comparisons of parametric tests, permutation tests based on original data, and permutation tests based on rarefied data.

3.7 Conclusion

We are not advocating the use of rarefying in all microbiome data analysis. Rather, we highlight that in certain scenarios rarefying can be a valuable tool to guarantee the validity of permutation tests, and the loss in sensitivity due to rarefying might not be as severe as one might imagine. If loss in power is of concern, sample REIs can be used as a heuristic to determine whether one should proceed with rarefaction. Whether to rarefy or not ultimately depends on assumptions of the data generating process and characteristics of the data.

Chapter 4

Assessing the association between library sizes and microbial compositions in microbiome studies

This chapter is based on joint work with Ulas Karaoz, William Fithian, and Perry de Valpine.

4.1 Introduction

High-throughput microbiome data samples often have widely varying numbers of sequences (known as *library sizes*). Library sizes variations are often believed to be an artifact of the sequencing procedure, for instance due to preferential amplification by polymerase chain reaction (PCR) [1]. As mentioned in [99], "the microbial community in each biological sample may be represented by very different numbers of sequences (i.e., library sizes), reflecting differential efficiency of the sequencing process rather than true biological variation".

To facilitate proper comparisons of microbial communities, various normalization strategies for microbiome data have been proposed. One particular approach is rarefaction, also referred as *rarefying* in [64], a normalization procedure first proposed in [83] to compare alpha diversity in oceanography. The intuitive motivation for rarefaction is that observations become "comparable" after rarefaction since rarefied data all share the same library size.

In the previous chapter, we developed a nonparametric graphical model for microbiome data to investigate the statistical trade-offs of rarefaction. Under the graphical model framework, we showed that rarefaction guarantees the validity of permutation tests for grouped microbiome data even if library sizes might be associated with the grouping variable of interest, and the loss in sensitivity due to rarefaction depends on the latent variation of microbial compositions. We also provided examples of how permutation tests can potentially be invalid without rarefying the data.

To formalize our discussion, we make a distinction between *latent microbial composition* and *observed microbial composition.* In microbiome data analysis, the quantity of scientific interest is the latent microbial composition, which can be thought of as the microbial composition of the sample after extraction and PCR. The adjective *latent* is used to emphasize that in practice we do not observe every individual sequence in a sample via sequencing so the microbial composition is never directly observed. On the other hand, the observed microbial composition, defined as the observed OTU (operational taxonomic unit) counts divided by the observed library size (this is often called total-sum scaling, TSS, in the literature), clearly depends on the library size. Conditional on the library size, the variance of the observed microbial composition decreases as the library size increases.

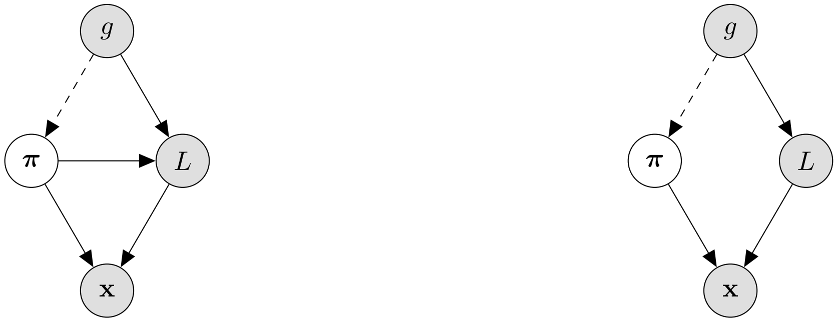
If library sizes were truly an artifact of the sequencing procedure, there would not be any association between the latent microbial composition and the library size. A natural question is how to assess this association from microbiome data. While the observed composition is a reasonable proxy for the latent composition, the observed composition always depends on the library size, so disentangling this dependence from the association between the latent composition and the library size can be challenging.

Another natural question is whether the association between the latent composition and the library size affects statistical comparisons of latent composition across different groups of observations. Broadly speaking, there are two types of inference, which we term *conditional inference* and *unconditional inference*. In conditional inference, statistical comparisons in latent comparison are performed conditioning on the library sizes; in unconditional inference statistical comparisons are made without the conditioning. For example, metagenomeSeq [69], a popular differential abundance testing method for microbiome data, allows both conditional inference and unconditional inference by having the library sizes as an optional argument in its software package [70].

In this work, under the nonparametric graphical model framework for microbiome data, we discuss how conditional inference and unconditional inference can lead to different conclusions about microbial compositions. In addition, we establish a sufficient condition, namely the conditional independence between the latent microbial composition and the library size given the grouping variable of interest, for conditional inference to be valid for testing the unconditional independence between the microbial composition and the grouping variable. Furthermore, we develop a rarefaction-based nonparametric statistical testing procedure, the *combined correlation permutation test*, to assess whether library sizes are associated with microbial compositions conditioning on the grouping variable of interest. We also discuss how multiple testing can be used to detect OTU-specific associations with library sizes. We apply these testing procedures to various microbiome datasets and find that such association arises quite often in practice.

4.2 Notations and model setup

Let G be the number of groups. For $g = 1, ..., G$, let n_g be the number of observations in group g. Denote the smallest group sizes as $n_{\min} = \min(n_1, ..., n_G)$ and the sum of the groups



(a) Full model: L depends on both $\boldsymbol{\pi}$ and g. (b) Simplified model: L depends on only g.

Figure 4.1: A non-parametric graphical model for grouped microbiome data. The shaded nodes are observed quantities. The dashed arrow represents the hypothesis of interest: whether microbial compositions vary across different groups.

sizes as N. For sample i in group g, let $L_i^{(g)}$ be the associated library size and $\mathbf{x}_i^{(g)}$ be the associated count vector of OTUs.

To simplify the setup, suppose we have two groups of n observations of count vector (the discussion can be generalized to more than two groups as well as unequal group sizes). For $i \in \{1, ..., n\}$ and $g \in \{1, 2\}$, denote the ith observation in group g as $x_i^{(g)} \in \mathbb{Z}_{\geq 0}^p$, where p is the number of species. Let $\pi_i^{(g)}$ be the latent composition and $L_i^{(g)} = \sum_{j=1}^p (x_i^{(g)})_j$ be the library size.

We assume the data is generated according to the following nonparametric graphical model:

$$\begin{aligned} &\text{latent composition:} & \boldsymbol{\pi}_i^{(g)} | g &\sim f_{\boldsymbol{\pi}}(\cdot | g) & (4.1) \\ &\text{library size:} & L_i^{(g)} | \boldsymbol{\pi}_i^{(g)}, g &\sim f_L(\cdot | \boldsymbol{\pi}_i^{(g)}, g) & (4.2) \\ &\text{count:} & \mathbf{x}_i^{(g)} | \boldsymbol{\pi}_i^{(g)}, L_i^{(g)} &\sim \text{Multinomial}(L_i^{(g)}, \boldsymbol{\pi}_i^{(g)}), & (4.3) \end{aligned}$$

where $f_{\boldsymbol{\pi}}(\cdot|g)$ is a probability density function supported on the p-dimensional probability simplex and $f_L(\cdot|\boldsymbol{\pi}_i^{(g)}, g)$ is a probability mass function supported on non-negative integers. We call this model the *full model* (Figure 4.1a). If we further assume the library size is independent of the latent composition (given the group membership), we call the model the the *simplified model* (Figure 4.1b).

The typical hypothesis of interest in microbiome studies is whether the distribution of the latent composition $\boldsymbol{\pi}$ is independent of g; that is $f_{\boldsymbol{\pi}}(\cdot|g) = f_{\boldsymbol{\pi}}(\cdot)$. In terms of the graphical model language, we are interested in whether there is an arrow from g to $\boldsymbol{\pi}$. We refer this hypothesis as the *unconditional null hypothesis*, since we are studying the independence between the latent composition and the grouping variable without conditioning on the library size.

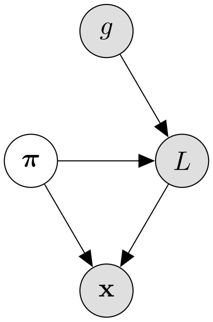


Figure 4.2: Full model with the unconditional null hypothesis: the microbial composition $\boldsymbol{\pi}$ is independent of the grouping variable g.

4.3 Conditional inference versus unconditional inference

We formalize under the nonparametric graphical model framework the difference between conditonal inference and unconditional inference.

Assume the full model holds with the unconditional null hypothesis: the microbial composition $\boldsymbol{\pi}$ is independent of the grouping variable g (Figure 4.2). Due to the explaining-away phenomenon, $\boldsymbol{\pi}$ and g are conditionally dependent given the library size L. This implies that the unconditional null hypothesis and the conditional null hypothesis are not equivalent under the full model, meaning that we can potentially arrive at different conclusions with conditional inference and unconditional inference.

For a concrete hypothetical example, suppose we have only $p = 2$ OTUs of interest and there are two groups of n observations. Suppose further the distribution of $\boldsymbol{\pi}$ does not depend on the group membership g (that is, the unconditional null hypothesis holds), and

$$\boldsymbol{\pi} = \begin{cases} (1/4, 3/4) & \text{with probability } 1/2 \\ (3/4, 1/4) & \text{with probability } 1/2 \end{cases}. \tag{4.4}$$

For some reason, the machine processing samples in group 1 always returns observations with library size 1000 whenever the first OTU is dominant and 2000 otherwise; the machine processing samples in group 2 always returns observations with library size 2000 whenever the first OTU is dominant and 1000 otherwise. Suppose we are interesting in the unconditional null hypothesis but we proceed with conditional inference. Since the unconditional null hypothesis holds, we commit a Type I error (false positive) if we reject the null hypothesis in subsequent testing procedures.

Since library sizes can only take two values in this example, conditioning on the library sizes is equivalent to creating two strata of the samples: stratum 1 contains all the observations with library size 1000, and stratum 2 contains all the observations with library size 2000. Now note that within stratum 1 $P(\boldsymbol{\pi} = (1/4, 3/4)|L = 1000, g = 1) = 1$ and $P(\boldsymbol{\pi} = (1/4, 3/4)|L = 1000, g = 2) = 0$. Therefore, conditioning on the library size any sensible statistical procedure (for example, running a two-sample t-test for the relative abundance of OTU 1 for each stratum) would conclude there is a difference in microbial compositions between the two groups of samples, commiting a Type I error with respect to the unconditional null hypothesis.

A natural question is under what scenarios conditional inference is valid for testing the unconditional null hypothesis. A sufficient condition is that the simplified model (Figure 4.1b) holds. In the simplified model, the latent composition $\boldsymbol{\pi}$ is conditionally independent of the library size L given the group membership g. The following theorem shows that under the simplified model, the conditional null hypothesis is equivalent to the unconditional null hypothesis.

Theorem 1. *Under the simplified model, $\boldsymbol{\pi}$ is independent of g if and only if $\boldsymbol{\pi}$ is conditionally independent of g given the library size L.*

Proof. (Unconditional independence implies conditional independence): Assume $\boldsymbol{\pi}$ is independent of g. From the simplified model, the unconditional null also implies that $\boldsymbol{\pi}$ is independent of (L, g). Therefore, for all L and g,

$$p(\boldsymbol{\pi}|L, g) = p(\boldsymbol{\pi}) = p(\boldsymbol{\pi}|L).$$

(Conditional independence implies unconditional independence): Assume $\boldsymbol{\pi}$ is conditionally independent of g given L. Then for all L and g,

$$p(\boldsymbol{\pi}|g) = p(\boldsymbol{\pi}|L, g) = p(\boldsymbol{\pi}|L), \tag{4.5}$$

where the first equality uses the conditional null hypothesis and the second equality uses the property of the simplified model. Let P_L denote the marginal distribution of L. From (4.5), for all g,

$$p(\boldsymbol{\pi}|g) = \int p(\boldsymbol{\pi}|g) dP_L(L) = \int p(\boldsymbol{\pi}|L) dP_L(L) = p(\boldsymbol{\pi}),$$

so π is indeed independent of g. □

4.4 Methodology

In this section we develop a hypothesis testing procedure for testing whether $\boldsymbol{\pi}$ and L are conditionally independent given the grouping variable g. Equivalently, this is testing whether the simplified model is adequate in modeling the data relative to the full model (Figure 4.3).

Within each group, to measure the dependence between the latent composition $\boldsymbol{\pi}$ and the library size L (here we drop the superscript (g) to simplify notations), we study the OTU-specific Spearman's correlation coefficients [88] between the latent proportions $\boldsymbol{\pi}_j$ and the library size L for each OTU $j = 1, ..., J$, where the Spearman's correlation coefficient between two variables is defined as the correlation between the ranked version of the two

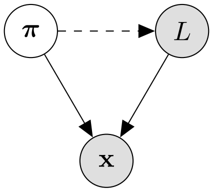


Figure 4.3: A graphical model representation of testing the simplified model against the full model for a particular group of observations. To simplify the representation, the node for the grouping variable g is dropped since we condition on g. The dashed arrow represents the hypothesis of whether microbial compositions and library sizes are associated. Inference on the dashed arrow based on $\mathbf{x}$ is challenging because of the direct dependence between $\mathbf{x}$ and L.

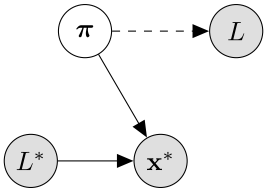


Figure 4.4: A graphical model representation of rarefied data. The rarefied depth is denoted by L^* and the corresponding rarefied data is denoted by $\mathbf{x}^*$. Since the only way rarefied count can depend on the original library size L is through the latent composition $\boldsymbol{\pi}$, it is sensible to use $\mathbf{x}^*$ to infer whether L and $\boldsymbol{\pi}$ are dependent.

variables. Spearman's correlation coefficient is used instead of Pearson's correlation coefficient because outlying library sizes are common in practice and Pearson's correlation coefficient is sensitive to outliers.

Since $\boldsymbol{\pi}$ is not observable, we cannot directly compute the sample correlation between π_j and L. A naive approach is to compute the sample Spearman's correlation coefficients between the observed relative abundances and the associated library sizes instead. However, conditional on the library sizes, under the simplified model the observed relative abundances are not exchangeable since the library size controls the variation of observed relative abundances. The lack of exchangeability is problematic because a randomization test on correlation might not be valid. Based on a result in the previous chapter, rarefied observations (to the same pre-specified rarefied depth L^*) from the same group are exchangeable. It is hence sensible to use the sample Spearman's correlation coefficients between the rarefied relative abundances and the associated library sizes. Rarefaction can be summarized as follows:

1. Select a rarefied depth L^*.

2. Discard all the samples with library size less than L^*.

3. Subsample all the samples with library size L_i greater than L^* to L^*. This subsampling is done via sampling without replacement.

Let $\mathbf{R}$ be the matrix of rarefied relative abundances and $\mathbf{R}_j$ be the vector of rarefied relative abundances for OTU j. Let $\mathbf{L}$ be the vector of original library sizes. For OTU $j = 1, ..., J$, let $T_j(\mathbf{R}, \mathbf{L}) = |\hat{\rho}(\mathbf{R}_j, \mathbf{L})|$ be the absolute value of the sample Spearman's correlation coefficients between the rarefied relative abundances for OTU j and the associated library sizes.

We keep the matrix $\mathbf{R}$ fixed and permute the vector of library sizes $\mathbf{L}$. Let $\mathbf{L}^{(1)}, ..., \mathbf{L}^{(B)}$ be B randomly permuted vectors of library sizes. Then for $b = 1, ..., B$, $T_j(\mathbf{R}, \mathbf{L}^{(b)})$ is the bth permuted test statistic for the jth OTU. Let $\#A$ denote the number of elements in the set A. For $j \in \{1, ..., J\}$, define the permutation p-value for the jth OTU as

$$p_j = \frac{1 + \#\{b : T_j(\mathbf{R}, \mathbf{L}^{(b)}) \geq T_j(\mathbf{R}, \mathbf{L})\}}{1 + B}, \tag{4.6}$$

which is the proportion of permuted test statistics (including the observed one) being at least as large as the observed test statistic. For our purpose of detecting the presence of *any* associations of compositions and library sizes, rather than testing for each individual OTU separately, we combine the p-values via Fisher's method [30] to form a combined statistic

$$Y = -2 \sum_{j=1}^{J} \log p_j, \tag{4.7}$$

which synthesizes the statistical evidence across all the OTUs. If the permutation p-values are approximately independent under the assumption that $\boldsymbol{\pi}$ and L are conditionally independent given g, we can simply use the χ^2 distribution with J degrees of freedom to compute the p-value associated with Fisher's combined statistic Y. However, since there are complex relationships among the OTUs, such an independent assumption is usually unwarranted.

To proceed, we obtain the approximate permutation distribution of Y. A direct approach to simulate the permutation distribution of Y is to generate a large number of permuted library size vectors, break them down into many smaller subsets, compute the combined statistic for each subset, and aggregate these statistics to get the approximate permutation distribution of Y. To save computations, we recycle the permuted library size vectors $\mathbf{L}^{(1)}, ..., \mathbf{L}^{(B)}$ to compute the bth permuted Fisher's combined statistic for $b = 1, ..., B$:

$$Y^{(b)} = -2 \sum_{j=1}^{J} \log \left[\frac{I(T_j(\mathbf{R}, \mathbf{L}) \geq T_j(\mathbf{R}, \mathbf{L}^{(b)})) + \#\{c : T_j(\mathbf{R}, \mathbf{L}^{(c)}) \geq T_j(\mathbf{R}, \mathbf{L}^{(b)})\}}{1 + B} \right]. \tag{4.8}$$

Since $(\mathbf{L}, \mathbf{L}^{(1)}, ..., \mathbf{L}^{(B)})$ is exchangeable, $(Y, Y^{(1)}, ..., Y^{(b)})$ is also exchangeable, so a valid permutation p-value can be constructed using $(Y, Y^{(1)}, ..., Y^{(b)})$. The permutation p-value based on Fisher's combined statistic is

$$p = \frac{1 + \#\{b : Y^{(b)} \geq Y\}}{1 + B}, \tag{4.9}$$

which is the proportion of permuted Fisher's combined statistics (including the observed one) being at least as large as the observed Fisher's combined statistic. If the permutation p-value is below a pre-specified significance level (say 5%), we conclude that $\boldsymbol{\pi}$ and L are conditionally dependent given g. We call this test the *combined correlation permutation test*. In the combined correlation permutation test, we obtain a single p-value based on one rarefied dataset. If one is concerned about using only one particular rarefied dataset, one can average the resulting p-values over different versions (say 10 different versions) of the rarefied datasets to lower the variance of the final p-value.

Recall that our goal is to test the independence between $\boldsymbol{\pi}$ and L within each group of observations. After obtaining the permutation p-value from combined correlation permutation test for each group, given a pre-specified significance level, we can apply the Bonferroni correction to determine there is sufficient statistical evidence for the $\boldsymbol{\pi}$-L conditional dependence. As long as there is at least one significant p-value, we conclude that there are $\boldsymbol{\pi}$-L conditional dependence.

We can also test for the conditional independence between the microbial proportion and the library size each of the particular OTU by computing a permutation p-value based on Spearman's correlation coefficient. This leads to a multiple testing problem, which can be tackled by applying a Bonferroni correction or the Benjamini-Hochberg procedure.

4.5 Case studies

We provide several real data examples to show how to apply the methodology in practice. In each of the case studies, for each group we use 200 permutations in the permutation tests (so the smallest possible permutation p-value is 0.005) and average the final p-value over 10 different rarefied datasets. We assume that the unconditional null hypothesis is of interest. We also study the OTU-specific conditional independence. We increase the number of permutations to the number of OTUs multiplied by 1000 to ensure that the resulting p-values can be small enough to reject the null hypothesis.

4.5.1 Human Microbiome Project

The first data example is from the Human Microbiome Project dataset (HMPv35, [42]), available in the R package MicrobesDS. The goal of the study is to determine if there is a core microbiome at each body site. Body sites include the gastrointestinal and female urogenital tracts, oral cavity, nasal and pharyngeal tract, and skin. For the purpose of illustration, we look at only tongue and throat. Suppose our scientific objective is to determine whether the tongue microbiome is different from the throat microbiome, without conditioning on the library sizes. The p-values for the combined correlation permutation test are 0.005 and 0.005 for tongue and throat respectively. Since there is strong statistical evidence that the simplified model is inadequate relative to the full model, conditional inference is inappropriate. Using the Bonferroni correction, only 0.6% of OTUs in tongue microbiomes and 0.1% of

OTUs in throat microbiomes shows significant dependence with library sizes; using the Benjamini-Hochberg procedure, 2% and 0.2% respectively.

4.5.2 Diets and infant microbiomes

The second data example is from an infant fecal microbiome study [36]. The goal is to investigate the impact of three different diets (breast milk, experimental infant formula, and standard infant formula) on infant fecal microbiomes via a randomized controlled trial. For the purpose of illustration we use only the data for six-month-old infants. The p-values from the combined correlation permutation test for the three diets are 0.036, 0.854 and 0.747 respectively. If we choose 5% as the significance level and apply Bonferroni's correction, none of the p-values are significant. The simplified model may be adequate to describe the data and conditional inference may be appropriate. For the OTU-specific test for conditional independence, we do not find any statistically significant result using either the Bonferroni correction or the Benjamini-Hochberg procedure.

4.5.3 Hand surface bacteria

The third data example is from a study on hand surface bacteria [29]. One of the objectives is to assess whether microbial composition is associated with time since last hand washing. We use time since last hand washing as the grouping variable, and restrict our attention on observations with time since last hand washing being two hours and six hours. The p-values are 0.123 and 0.044 respectively. If we choose 5% as the significance level and apply Bonferroni's correction, none of the p-values are significant. Similar to the previous case studies, the simplified model may be plausible. For the OTU-specific test for conditional independence, we do not find any statistically significant result using either the Bonferroni correction or the Benjamini-Hochberg procedure.

4.5.4 Air of bedrooms in the Chicago area

The fourth data example is from a study on microbiome and allergens in the air of bedroom in the Chicago area [76]. We use neighborhood as the grouping variable, and use only observations in suburban neighborhoods as well as observations in urban neighborhoods. The p-values are 0.005 and 1.000 respectively. The small p-value for suburban neighborhoods provides evidence that the simplified model is not appropriate for the data, so conditional inference is inappropriate. For the OTU-specific test for conditional independence, we do not find any statistically significant result using either the Bonferroni correction or the Benjamini-Hochberg procedure.

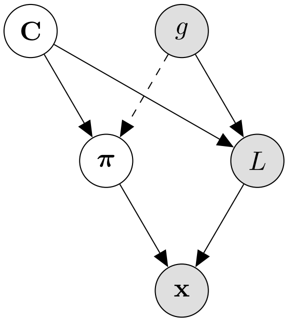


Figure 4.5: The association between the latent composition $\boldsymbol{\pi}$ and the library size L can arise if there are unobserved variables $\mathbf{C}$ affecting both $\boldsymbol{\pi}$ and L.

4.5.5 Forensic identification

The fifth data example is from a forensic identification study [28]. One of the research objectives is to study the microbial compositional differences between skin surfaces and keyboard surfaces. The p-values for the two groups are 0.005 and 0.005 respectively. Such small p-values indicate strong statistical evidence for the dependence between microbial compositions and library sizes, and conditional inference is not appropriate. For the OTU-specific test for conditional independence, we do not find any statistically significant result using either the Bonferroni correction or the Benjamini-Hochberg procedure.

4.6 Discussion and conclusion

From the case studies, library sizes seem to be often associated with latent compositions, challenging the conventional wisdom that library sizes are simply artifacts of the sequencing procedures. While there might not be a direct relationship between the latent composition $\boldsymbol{\pi}$ and the library size L, associations between these two quantities can arise from unobserved variables. Consider the graphical model in Figure 4.5. The unobserved variables $\mathbf{C}$ affect both $\boldsymbol{\pi}$ and L but $\mathbf{C}$ do not depend on the grouping variable g. Even if $\boldsymbol{\pi}$ is independent of g, $\boldsymbol{\pi}$ and g becomes dependent through the path $g \rightarrow L \leftarrow \mathbf{C} \rightarrow \boldsymbol{\pi}$ conditioning on L.

If the unconditional null hypothesis is of scientific interest, one can use the combined correlation permutation test as a preliminary indicator of the invalidity of conditional inference. Suppose we choose a cutoff, say 5%, for the p-values. If the test yields at least one p-value less than 5%, there is statistical evidence that simplified model does not hold and conditional inference might not be valid. On the other hand, if all the p-values are larger than 5%, there is no evidence that simplified model does not hold but we cannot simply conclude that simplified model holds without further assumptions. It is possible that there are other types of conditional dependence not well captured by Spearman's correlation coefficients, or the test is not sensitive enough to yield small p-values. This is analogous to how one

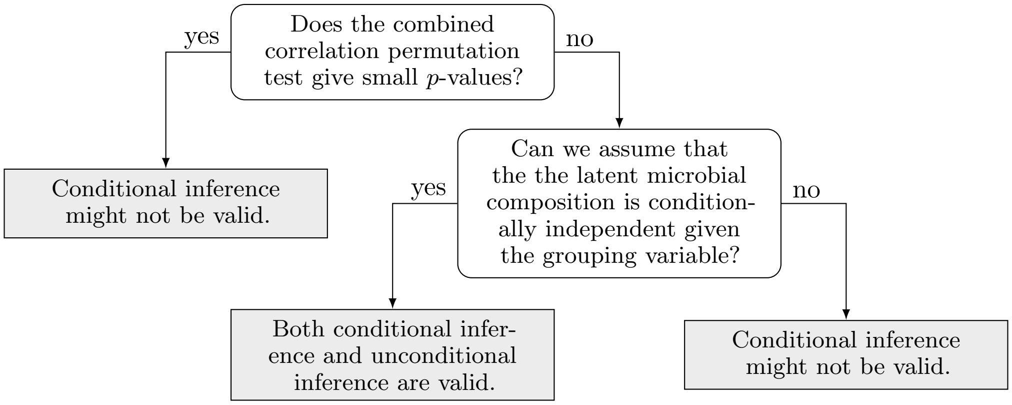

Does the combined correlation permutation test give small p-values?
yes
no
Conditional inference might not be valid.
Can we assume that the the latent microbial composition is conditionally independent given the grouping variable?
yes
no
Both conditional inference and unconditional inference are valid.
Conditional inference might not be valid.

Figure 4.6: A workflow to determine whether conditional inference is valid for testing the unconditional null hypothesis: microbial composition is independent of the grouping variable.

cannot conclude good model fit from failing to reject a goodness-of-fit test. The workflow to determine whether conditional inference is valid for testing the unconditional null hypothesis is summarized in Figure 4.6.

For the OTU-specific test for conditional independence, one potential reason for the lack of significant results is that permutation p-values cannot be too small for OTUs with many zeros. After a multiple testing correction, many of the OTU-specific permutation p-values become insignificant.

If a researcher is concerned about the validity of conditional inference and is not comfortable making further assumptions on the data generating process, one can first rarefy the data and apply the conditional inference methods. This yields valid inference because rarefaction removes the dependence between the library sizes and the latent composition (as well as the grouping variable).

Acknowledgements

We thank Christina Jin for preprocessing microbiome datasets from Qiita.

Chapter 5

Future directions

This dissertation has explored both computational and applied aspects of latent variable models. In particular, we studied efficient maximum likelihood estimation methods for latent variable models and developed a non-parametric graphical model framework for microbiome data analysis. We discuss a few future directions in these lines of research.

5.1 Maximum likelihood estimation

With the use of “warm start” the algorithms might afford to have a much smaller burn-in, rather than half of the MCMC samples, since the starting point is reasonably close to the high density region of the stationary distribution. It would be interesting to formalize the benefits of “warm start” in the MCMC sampling part of the algorithms. We remark that a hybrid approach, such as running 1D sampling initially and then switching to MCNR or Adam, might be beneficial. However, determining when to switch to the other algorithm can be tricky to automate. Although the examples shown are on the smaller side of latent variable model problems, there is potential to consider leveraging tools such as *greta* [34], a modeling framework in R that uses Google’s TensorFlow, to scale our approach without sacrificing accessibility and usability for practitioners.

5.2 Stratified rarefaction: a data-efficient normalization for permutation inference in microbiome studies

The idea of stratified rarefaction is as follows:

1. Specify a protocol on how to create strata of observations based on library sizes. Each stratum must contain at least one observation from each group.

2. Within each stratum, subsample each of the observations to the smallest library size in the stratum.

One can discard all the observations with unreasonably small library sizes prior to the stratified rarefaction procedure. An example of a stratification protocol is to pair up the observations with the same library size rank. That is, create n strata such that the kth stratum consists of the kth smallest (in terms of library size) observation from each group. Another example is to partition the positive real line into intervals; for each interval create a stratum by collecting all the observations with library sizes inside the interval.

After stratified rarefaction, one can carry out a conditional permutation test; we permute group labels of the observations within each stratum, not across different strata. The conditional permutation test is valid for the null hypothesis that the latent composition π is conditionally independent of the group membership g given the library size L. We call this the *conditional null hypothesis*. This is different from the *unconditional null hypothesis*: the latent composition π is independent of the group membership g. Both hypotheses could be of scientific interest, although the unconditional null hypothesis is more natural if library sizes. As discussed in Chapter 4, these two hypotheses are not equivalent under the full model but they are equivalent under the simplified model.

5.2.1 Stratum specification

Ideally the stratification scheme should be chosen to maximize the power of the subsequent hypothesis testing procedures. However, power calculations are dependent on the choice of the tests and generally quite involved. A more straightforward objective is to minimize the total loss in library sizes. We can formulate this as an optimization problem. Define $S+1$ cutoff points $a_0, a_1, ..., a_S$ with $a_0 = \min(L_i^{(g)})$, $a_S = \infty$, and $a_{s-1} \leq a_s$ for $s = 1, ..., S$ to discretize the range of library sizes. Given that our goal is to minimize library size loss, we would like to solve the following optimization problem

$$\min_{S} \min_{a_1,\ldots,a_{S-1}} \sum_{g=1}^{G} \sum_{i=1}^{n_g} \sum_{s=1}^{S} (L_i^{(g)} - a_{s-1}) I(a_{s-1} \leq L_i^{(g)} < a_s) \tag{5.1}$$

subject to the constraint that each interval $[a_{s-1}, a_s)$ must contain at least one library size from each group. Due to the combinatorial nature of the optimization problem, finding the optimal solution is not computationally feasible even for moderately sized problems.

An alternative approach is to assign ranks for library sizes within each group of observations, and to stratify observations by matching observations across groups based on their ranks. We call this type of stratified rarefaction *pairwise rarefaction*.

5.2.2 Number of permutations

By restricting the set of possible permutations, a natural concern is whether there are sufficient permutations for small p-values. This is especially concerning if multiple testing

	$G = 2$		$G = 3$	
n	rarefying	pairwise rarefying	rarefying	pairwise rarefying
5	3.63×10^{6}	32	1.31×10^{12}	7776
10	2.43×10^{18}	1024	2.65×10^{32}	6.05×10^{7}
15	2.65×10^{32}	32768	1.20×10^{56}	4.70×10^{11}
20	8.16×10^{47}	1.05×10^{6}	8.32×10^{81}	3.66×10^{15}

	$G = 4$		$G = 5$	
n	rarefying	pairwise rarefying	rarefying	pairwise rarefying
5	2.43×10^{18}	7.96×10^{6}	1.55×10^{25}	2.49×10^{10}
10	8.16×10^{47}	6.34×10^{13}	3.04×10^{64}	6.19×10^{20}
15	8.32×10^{81}	5.05×10^{20}	2.48×10^{109}	1.54×10^{31}
20	7.16×10^{118}	4.02×10^{27}	9.33×10^{157}	3.83×10^{41}

Table 5.1: Number of permutations for rarefying and pairwise rarefying for comparing G groups ($G \in \{2, 3, 4, 5\}$) with the same number of observations n.

corrections are being applied. In a permutation test, the smallest possible p-value is 1/#Perm, where #Perm is the number of distinct permutations. Suppose we partition the data into S strata. For $s = 1, ..., S$, let $n_{g,s}$ be the number of observations in group g in stratum s and $n_{\cdot,s}$ be the number of observations in stratum s. We require $n_{g,s} \geq 1$ for all g and s to ensure that each stratum has at least one observation from each group. The number of permutations for stratified rarefying is $\prod_{s=1}^{S} \left(n_{\cdot,s}!\right)$. In particular, if N is the total number of observations, for rarefaction, the number of permutations is $N!$; for pairwise rarefaction, $G!^{n_{\min}-1}[N - G(n_{\min} - 1)]!$, where $n_{\min} = \min(n_1, ..., n_G)$. Table 5.1 displays the number of permutations for rarefaction and pairwise rarefaction when there are $G \in \{2, 3, 4, 5\}$ groups of equal number of observations. As a rough guideline for a reasonable smallest permutation p-value, for $G = 2$, we recommend at least 15 observations in each group for pairwise rarefaction; for $G = 3$, at least 10. For $G \geq 4$, the number of permutations for pairwise rarefaction is large even when the sample size n is only 5.

5.3 Dependence between the library size and the grouping variable in microbiome data analysis

Chapter 4 of this dissertation have developed methodology for studying the association between the latent composition and the library size and have found that such dependence are not uncommon in practice. As discussed in Chapter 4, this association might undermine the use of conditional inference on the unconditional null hypothesis. On the other hand, in the nonparametric graphical model, there is an arrow from the grouping variable g to the library size L. Without this arrow, observations are exchangeable even if data have not been rarefied, implying that permutation inference is valid with original data. Hence it is of interest to determine whether this association between the grouping variable and the library size is common in practice.

There has been ongoing effort in this research direction with Christina Jin, William Fithian, Perry de Valpine, and Ulas Karaoz. Preliminary meta-analysis based on about 20 microbiome datasets from the open-source microbial study management platform Qiita[1] has shown that it is indeed not uncommon that library size distributions depend on the grouping variable. Such phenomenon remains prominent even when more sophisticated library size estimation methods, cumulative-sum-scaling (CSS [69]) and geometric mean of pairwise ratios (GMPR [16]), are used.

Another research direction is to study how much library size distributions have to differ across groups in order for permutation inference to be invalid. As shown in Section **??**, it is possible that permutation inference remains valid even when the exchangeability assumption in permutation test is not met. There are several challenges in this research direction. First, the robustness to difference in library distributions is likely to depend on the particular test statistic used. Second, it is unclear which particular aspects of library size distributions can undermine permutation inference. Potential aspects to be considered include how large the difference in mean library sizes is, how large the difference in spreads is, and how much the library size distributions overlap.

[1] `https://qiita.ucsd.edu/`.

Bibliography

[1] D. Aird et al. "Analyzing and minimizing PCR amplification bias in illumina sequencing libraries". In: *Genome Biology* 12.2 (2011).

[2] M. J. Anderson. "A new method for non-parametric multivariate analysis of variance". In: *Austral ecology* (2001), pp. 32–46.

[3] S. Asmussen and P. W. Glynn. *Stochastic Simulation: Algorithms and Analysis*. Stochastic Modelling and Applied Probability. Springer, 2007.

[4] D. Bates et al. "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 67.1 (2015), pp. 1–48.

[5] M. J. Beal and Z. Ghahramani. "The Variational Bayesian EM Algorithm for Incomplete Data: with Application to Scoring Graphiclal Model Structures". In: *Bayesian Statistics* 7 (2003).

[6] L. Bottou, ed. *Large-Scale Machine Learning with Stochastic Gradient Descent*. COMPSTAT. 2010.

[7] L. Bottou, ed. *Stochastic gradient learning in neural networks*. Neuro-Nimes. 1991.

[8] S. Brooks et al. *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2011.

[9] B. S. Caffo, W. Jank, and G.L. Jones. "Ascent-based Monte Carlo expectation maximization". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2005), pp. 235–251.

[10] O. Cappé, E. Moulines, and T. Ryden. *Inference in Hidden Markov Models*. 1. Springer-Verlag New York, 2005.

[11] J.-F. Cardoso, M. Lavielle, and E. Moulines. "Un algorithme d'identification par maximum de vraisemblance pour des données incomplètes". In: *C.R. Acad. Sci. Paris Série I Statistique* 320 (1995), pp. 363–368.

[12] B. Carpenter et al. "Stan: A probabilistic programming language". In: *Journal of Statistical Software* 76.1 (2017).

[13] G. Celeux and J. Diebolt. "The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem." In: *Computational Statistics* 2 (1985), pp. 73–82.

[14] G. Celeux and J. Diebolt. “Une version de type recuit simulé de l’algorithme EM.” In: *C. R. Acad.Sci. Paris Sér. I Math.* 310 (1990), pp. 119–124.

[15] J. Chen and H. Li. “Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis”. In: *Annals of Applied Statistics* 7.1 (2013), pp. 418–442.

[16] L. Chen et al. “GMPR: A robust normalization method for zero-inflated count data with application to microbiome sequencing data”. In: *PeerJ* 6.e4600 (2018).

[17] Y. B. Cheung et al. “Gut microbiota in Malawian infants in a nutritional supplementation trial”. In: *Tropical Medicine and International Health* 21.2 (2015), pp. 283–290.

[18] S. Chib and I. Jeliazkov. “Marginal likelihood from the Metropolis–Hastings output.” In: *Journal of the American Statistical Association* 96.453 (2001), pp. 270–281.

[19] J.S. Clark. “Why environmental scientists are becoming Bayesians.” In: *Ecology letters* 8.1 (2005), pp. 2–14.

[20] N. Cressie et al. “Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling.” In: *Ecological Applications* 19.3 (2009), pp. 553–570.

[21] M. J. Crowder. “Beta-binomial Anova for proportions”. In: *Journal of the Royal Statistical Society (C, Applied Statistics)* 27.1 (1978), pp. 34–37.

[22] P. Das and S. Ghosal. “Bayesian non-parametric simultaneous quantile regression for complete and grid data”. In: *Computational Statistics & Data Analysis* 127 (2018).

[23] B. Delyon, M. Lavielle, and E. Moulines. “Convergence of a stochastic approximation version of the EM algorithm”. In: *The Annals of Statistics* 27.1 (1999), pp. 94–128.

[24] A. P. Dempster, N. M. Laird, and D. B. Rubin. “Maximum Likelihood from Incomplete Data via the EM Algorithm”. In: *Journal of the Royal Statistical Society: Series B (methological)* (1977), pp. 1–38.

[25] J. Duchi, E. Hazan, and Y. Singer. “Adaptive Subgradient Methods for Online Leaning and Stochastic Optimization”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2121–2159.

[26] B. Efron. “Bayesians, frequentists, and scientists”. In: *Journal of the American Statistical Association* 100.469 (2005), pp. 1–5.

[27] B. Efron. “Discussion on Maximum Likelihood from Incomplete Data via the EM Algorithm”. In: *Journal of the Royal Statistical Society: Series B (methodological)* 39.29 (1977).

[28] N. Fierer et al. “Forensic identification using skin bacterial communities”. In: *Proceedings of the National Academy of Sciences* 107.14 (2010), pp. 6477–6481.

[29] N. Fierer et al. “The influence of sex, handedness, and washing on the diversity of hand surface bacteria”. In: *Proceedings of the National Academy of Sciences* 105.46 (2008), pp. 17994–17999.

[30] R. A. Fisher. *Statistical Methods for Research Workers*. 4th. Edinburgh: Oliver and Boyd, 1932.

[31] D. P. Gaver and I. G. O’Muircheartaigh. “Robust Empirical Bayes Analyses of Event Rates”. In: *Techometrics* 29.1 (1987), pp. 1–15.

[32] E. I. George, U. E. Makov, and A. F. M. Smith. “Conjugate Likelihood Distributions”. In: *Scandinavian Journal of Statistics* 20 (1993), pp. 147–156.

[33] C. J. Geyer and E. A. Thompson. “Constrained Monte Carlo Maximum Likelihood for Dependent Data”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 54.3 (1992), pp. 657–699.

[34] N. Golding. *greta: Simple and Scalable Statistical Modelling in R*. 2018. URL: `https://greta-stats.org/`.

[35] J. Halfvarson, C. Brislawn, R. Lamendella, et al. “Dynamics of the human gut microbiome in inflammatory bowel disease”. In: *Nat Microbiol* 2.17004 (2017).

[36] X. He et al. “Fecal microbiome and metabolome of infants fed bovine MFGM supplemented formula or standard formula with breast-fed infants as reference: a randomized controlled trial”. In: *Sci Rep* 9.11589 (2019).

[37] G. E. Hinton, S. Osindero, and Y. Teh. “A Fast Learning Algorithm For Deep Belief Networks”. In: *Neural Computation* 18 (2006), pp. 1527–1554.

[38] G.E. Hinton and R.R. Salakhutdinov. “Reducing the dimensionality of data with neural networks”. In: *Science* 313.5786 (2006), pp. 504–507.

[39] I. Holmes, K. Harris, and C. Quince. “Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics”. In: *PLoS One* 7 (2012).

[40] M.B. Hooten and N.T. Hobbs. “A guide to Bayesian model selection for ecologists”. In: *Ecological Monographs* 85.1 (2015), pp. 3–28.

[41] M. C. Horner-Devine et al. “A taxa-area relationship for bacteria”. In: *Nature* 432.7018 (2004), pp. 750–753.

[42] C. Huttenhower et al. “Structure, function and diversity of the healthy human microbiome”. In: *Nature* 486 (2012), pp. 207–214.

[43] S. Jackman. *Bayesian analysis for the social sciences.* Vol. 846. John Wiley and Sons, 2009.

[44] E. Jacquier, M. Johannes, and N. Polson. “MCMC maximum likelihood for latent state models”. In: *Journal of Econometrics* 137 (2007), pp. 615–640.

[45] S. Jangi, R. Gandhi, L. Cox, et al. “Alterations of the human gut microbiome in multiple sclerosis”. In: *Nat Commun* 7.12015 (2016).

[46] J. Jernvall and P. C. Wright. “Diversity components of impending primate extinctions”. In: *Proc Natl Acad Sci U S A* 95.19 (1998), pp. 11279–11283.

[47] M. R. Karim and S. L. Zeger. “Generalized Linear Models with Random Effects; Salamander Mating Revisited”. In: *Biometrics* (1992), pp. 631–644.

[48] B. Karimi, M. Lavielle, and E. Moulines. “f-SAEM: A fast Stochastic Approximation of the EM algorithm for nonlinear mixed effects models”. hal-01958248. 2018.

[49] R. L. Kashyap and C. C. Blaydon. “Estimation of Probability Density and Distribution Functions”. In: *IEEE Transactions on INformation Theory* (1968).

[50] J. Kiefer and J. Wolfowitz. “Stochastic Estimation of the Maximum of a Regression Function”. In: *The Annals of Mathematical Statistics* (1952).

[51] D. P. Kingma and J. L. Ba. “Adam: A Method for Stochastic Optimization”. In: *ICLR 2015* (2015).

[52] C. Knudson. *glmm: Generalized Linear Mixed Models via Monte Carlo Likelihood Approximation.* 2016. URL: `https://CRAN.R-project.org/package=glmm`.

[53] E. Kuhn and M. Lavielle. “Coupling a stochastic approximation version of EM with an MCMC procedure”. In: *ESAIM: Probability and Statistics* 8 (2004), pp. 115–131.

[54] A. Y. C. Kuk and Y. W. Cheng. “The Monte Carlo Newton-Raphson Algorithm”. In: *Journal of Statistical Computation and Simulation* 59 (1997), pp. 233–250.

[55] P. S. La Rosa et al. “Hypothesis Testing and Power Calculations for Taxonomic-Based Human Microbiome Data”. In: *PLOS ONE* 7.12 (2012).

[56] A.B. Lawson. *Bayesian disease mapping: hierarchical modeling in spatial epidemiology.* CRC press, 2013.

[57] S. R. Lele, B. Dennis, and F. Lutscher. “Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods”. In: *Ecology Letters* (2007).

[58] T. A. Louis. “Finding the Observed Information Matrix when Using the EM Algorithm”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 44.2 (1982), pp. 226–233.

[59] M. I. Love, W. Huber, and S. Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome Biology* 15.550 (2014).

[60] C. Lozupone and R. Knight. “UniFrac: a new phylogenetic method for comparing microbial communities”. In: *Applied and Environmental Microbiology* 71 (2005), pp. 8228–8235.

[61] C. Lozupone et al. “UniFrac: an effective distance metric for microbial community comparison”. In: *The ISME Journal* 5 (2011), pp. 169–172.

[62] D. Lunn et al. *The BUGS Book: A Practical Introduction to Bayesian Analysis.* CRC Press, 2012.

[63] C. E. McCulloch. "Maximum Likelihood Algorithms for Generalized Linear Mixed Models". In: *Journal of American Statistical Association* 92 (1997), pp. 162–170.

[64] P. J. McMurdie and S. Holmes. "Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible". In: *PLoS Comput Biol* 10.4 (2014).

[65] NIMBLE Development Team. *Automatic differentiation in NIMBLE*. 2019. URL: `https://github.com/nimble-dev/nimble/wiki/Automatic-differentiation-in-NIMBLE`.

[66] NIMBLE Development Team. *NIMBLE: An R Package for Programming with BUGS models, Version 0.6-3*. 2016. URL: `http://r-nimble.org`.

[67] O. Ovaskainen et al. "How to make more out of community data? A conceptual framework and its implementation as models and software". In: *Ecology Letters* 5 (2017), pp. 561–576.

[68] M. Parzen et al. "A generalized linear mixed model for longitudinal binary data with a marginal logit link function." In: *The annals of applied statistics* 5.1 (2011), pp. 449–467.

[69] J. N. Paulson et al. "Differential abundance analysis for microbial marker-gene surveys". In: *Nat Methods* 10 (2013), pp. 1200–1202.

[70] J. N. Paulson et al. "metagenomeSeq: Statistical analysis for sparse high-throughput sequncing". In: *Bioconductor package version 1.18.0 (Bioconductor)* (2017).

[71] B. Phipson et al. "Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression". In: *Annals of Applied Statistics* 10.2 (2016), pp. 946–963.

[72] U. Picchini and A. Samson. "Coupling stochastic EM and approximate Bayesian computation for parameter inference in state-space models". In: *Computational Statistics* 33.1 (2018), pp. 179–212.

[73] A. Plakhov and P. Cruz. "A Stochastic Approximation Algorithm with Step-Size Adaptation". In: *Journal of Mathematical Sciences* 120.1 (2004).

[74] Martyn Plummer. *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. 2003.

[75] A. Rakhlin, O. Shamir, and K. Sridharan. "Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization". In: *International Conference on Machine Learning* (2012).

[76] M. Richardson et al. "Concurrent measurement of microbiome and allergens in the air of bedrooms of allergy disease patients in the Chicago area". In: *Microbiome* 7.82 (2019).

[77] M. E. Ritchie et al. "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic Acids Research* 43.7 (2015).

[78] H. Robbins and S. Monro. “A Stochastic Approximation Method”. In: *The Annals of Mathematical Statistics* 22.3 (1951), pp. 400–407.

[79] M. D. Robinson, D.J. McCarthy, and G. K. Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.” In: *Bioinformatics* 26.1 (2010), pp. 139–140.

[80] T. M. K. Roeder et al., eds. *Warm starting Bayesian optimization.* Winter Simulation Conference. 2016.

[81] A.J. Royle and R.M. Dorazio. *Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities.* Academic Press, 2008.

[82] H. Rue, S. Martino, and N. Chopin. “Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations”. In: *Journal of the royal statistical society: Series b (statistical methodology)* 71.2 (2008), pp. 319–392.

[83] H. L. Sanders. “Marine benthic diversity: a comparative study”. In: *The American Naturalist* 102 (1968), pp. 243–282.

[84] B. W. Silverman. *Density estimation for statistics and data analysis.* Chapman and Hall, 1986.

[85] S. J. Song et al. “Preservation Methods Differ in Fecal Microbiome Stability, Affecting Suitability for Field Studies”. In: *mSystems* 1.3 (2016).

[86] J. C. Spall. “A Stochastic Approximation Technique for Generating Maximum Likelihood Parameter Estimates”. In: *Proceedings of American Control Conference* (1987).

[87] J.C. Spall. “Multivariate stochastic approximation using a simultaneous perturbation gradient approximation.” In: *IEEE Transactions on Automatic Control* (1992).

[88] C. Spearman. “The proof and measurement of association between two things”. In: *American Journal of Psychology* 15.1 (1904), pp. 72–101.

[89] A. W. van der Vaart. *Asymptotic Statistics.* Cambridge University Press, 1998.

[90] P. de Valpine. “Monte Carlo State-Space Likelihoods by Weighted Posterior Kernel Density Estimation”. In: *Journal of the American Statistical Association* 99 (2004), pp. 523–536.

[91] Perry de Valpine. “Frequentist analysis of hierarchical models for population dynamics and demographic data”. In: *Journal of Ornithology* 152 (2012), pp. 393–408. ISSN: 2193-7192. DOI: `10.1007/s10336-010-0642-5`.

[92] Perry de Valpine et al. “Programming With Models: Writing Statistical Algorithms for General Model Structures With NIMBLE”. In: *Journal of Computational and Graphical Statistics* (2017).

[93] D. Vandeputte, G. Kathagen, K. D’hoe, et al. “Quantitative microbiome profiling links gut community variation to microbial load”. In: *Nature* 551 (2017), pp. 507–511.

[94] N.M. Vogt, R.L. Kerby, K.A. Dill-McFarland, et al. "Gut microbiome alterations in Alzheimer's disease". In: *Sci Rep* 7.13537 (2017).

[95] M. J. Wainwright and M. I. Jordan. "Graphical models, exponential families, and variational inference". In: *Foundations and trends in machine learning* 1-2 (2008), pp. 1–305.

[96] A. Wald and J. Wolfowitz. "On a Test Whether Two Samples are from the Same Population". In: *The Annals of Mathematical Statistics* 11.2 (1940), pp. 147–162.

[97] K. Wang, T. Bui-Thanh, and O. Ghattas. "The Variational Bayesian EM Algorithm for Incomplete Data: with Application to Scoring Graphical Model Structures". In: *SIAM Journal on Scientific Computing* 40.1 (2018).

[98] G. C.G. Wei and M. A. Tanner. "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms". In: *Journal of the American Statistical Association* 85.411 (1990), pp. 699–704.

[99] S. Weiss et al. "Normalization and microbial differential abundance strategies depend upon data characteristics". In: *Microbiome* 5.27 (2017).

[100] J. R. White, N. Nagarajan, and M. Pop. "Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples". In: *PLOS Computational Biology* 5.4 (2009).

[101] A. D. Willis. "Rarefaction, Alpha Diversity, and Statistics". In: *Frontiers in Microbiology* 10.2407 (2019).

[102] P. Wolfe. "Convergence Conditions for Ascent Methods". In: *SIAM Review* 11.2 (1969), pp. 226–235.

[103] J. Wu, B. Peters, C. Dominianni, et al. "Cigarette smoking and the oral microbiome in a large study of American adults". In: *ISME J* 10 (2016), pp. 2435–2446.

[104] Y. Xia and J. Sun. "Hypothesis testing and statistical analysis of microbiome". In: *Genes and Diseases* 4.3 (2017), pp. 138–148.

[105] J. Xu, Y. Zhang, P. Zhang, et al. "The structure and function of the global citrus rhizosphere microbiome". In: *Nat Commun* 9.4894 (2018).

[106] M. D. Zeiler. "Adadelta: an adaptive learning method". arXiv:1212.5701. 2012.

[107] B. Zhou, L. Gao, and Y. Dai. "Gradient Methods with Adaptive Step-Sizes". In: *Computational Optiization and Applications* 35.1 (2006), pp. 69–86.

Appendix A

Supplementary materials for Chapter 2

In the following tables, Exec.(s) refers to the CPU time for 300 iterations in terms of seconds; Conv.(s) refers to the CPU time to convergence in terms of seconds; Conv.(iter.) refers to the number of iterations to convergence; loglik diff. refers to the log likelihood difference between the resulting estimate and the benchmark estimate; MSE refers to the mean-squared deviation between the resulting estimate and the benchmark estimate. NA in Conv. means that the convergence test is not passed within 300 iterations.

A.1 Numerical results for pump

The benchmark MLE $(0.823, 1.262)$ can be obtained numerically.

Table A.1: Numerical estimates for the *pump* example with MCMC sample size **300** and initial value $(\mathbf{10}, \mathbf{10})$.

	α	β
Fixed	0.863	1.365
Fixed 0.005	4.671	11.253
Adam	0.824	1.271
Newton-Raphson	0.820	1.252
1D Sampling	0.864	1.337
MCEM	0.825	1.267

Table A.2: Numerical performances for the *pump* example with MCMC sample size **300** and initial value $(\mathbf{10}, \mathbf{10})$.

	Exec.(s)	Conv.(s)	Conv.(iter.)	loglik diff.	MSE
Fixed	6.507	NA	NA	0.00827	0.00615
Fixed 0.005	6.115	NA	NA	9.90228	57.31417
Adam	6.418	3.534	167	0.00013	0.00005
Newton-Raphson	11.518	2.549	66	0.00007	0.00005
1D Sampling	17.604	3.765	67	0.00632	0.00364
MCEM	124.200	NA	NA	0.00002	0.00002

Table A.3: Numerical estimates for the *pump* example with MCMC sample size **300** and initial value $(\mathbf{10}, \mathbf{2})$.

	α	β
Fixed	0.820	1.254
Fixed 0.005	2.505	6.241
Adam	0.825	1.263
Newton-Raphson	0.817	1.248
1D Sampling	0.818	1.294
MCEM	0.825	1.267

Table A.4: Numerical performances for the *pump* example with MCMC sample size **300** and initial value $(\mathbf{10}, \mathbf{2})$.

	Exec.(s)	Conv.(s)	Conv.(iter.)	loglik diff.	MSE
Fixed	6.437	5.937	276	0.00005	0.00004
Fixed 0.005	6.305	NA	NA	4.50968	13.81102
Adam	6.516	NA	NA	0.00002	0.00000
Newton-Raphson	11.254	1.699	46	0.00016	0.00011
1D Sampling	16.203	4.908	90	0.00327	0.00053
MCEM	124.200	NA	NA	0.00002	0.00002

Table A.5: Numerical estimates for the *pump* example with MCMC sample size **3000** and initial value $(\mathbf{10}, \mathbf{10})$.

	α	β
Fixed	0.864	1.372
Fixed 0.005	4.676	11.252
Adam	0.817	1.248
Newton-Raphson	0.822	1.259
1D Sampling	0.855	1.304
MCEM	0.825	1.267

Table A.6: Numerical performances for the *pump* example with MCMC sample size **3000** and initial value $(\mathbf{10}, \mathbf{10})$.

	Exec.(s)	Conv.(s)	Conv.(iter.)	loglik diff.	MSE
Fixed	37.647	NA	NA	0.00922	0.00692
Fixed 0.005	34.470	NA	NA	9.90997	57.32270
Adam	37.415	19.806	162	0.00016	0.00011
Newton-Raphson	69.421	11.486	51	0.00000	0.00000
1D Sampling	49.248	9.918	62	0.00425	0.00142
MCEM	124.200	NA	NA	0.00002	0.00002

Table A.7: Numerical estimates for the *pump* example with MCMC sample size **3000** and initial value $(\mathbf{10}, \mathbf{2})$.

	α	β
Fixed	0.822	1.259
Fixed 0.005	2.505	6.236
Adam	0.824	1.265
Newton-Raphson	0.823	1.263
1D Sampling	0.851	1.326
MCEM	0.825	1.267

Table A.8: Numerical performances for the *pump* example with MCMC sample size **3000** and initial value $(\mathbf{10}, \mathbf{2})$.

	Exec.(s)	Conv.(s)	Conv.(iter.)	loglik diff.	MSE
Fixed	36.951	NA	NA	0.00001	0.00000
Fixed 0.005	35.305	NA	NA	4.50699	13.78914
Adam	37.211	36.962	298	0.00001	0.00001
Newton-Raphson	66.333	13.068	59	0.00000	0.00000
1D Sampling	48.362	11.841	74	0.00347	0.00242
MCEM	124.200	NA	NA	0.00002	0.00002

A.2 Numerical results for seeds

The *lme4* estimate (in bold) should be viewed as the benchmark for the MLE estimate. The glmer package relies heavily on the Gaussian assumption for random effects and leverages special case computations to drastically reduce the computational time.

Table A.9: Numerical estimates for the *seeds* example with MCMC sample size **20** and initial value $(\mathbf{0}, \mathbf{0}, \mathbf{1})$.

	β_0	β_1	σ_{RE}
Fixed 0.05	19.379	36.515	10.000
Fixed 0.005	5.177	5.005	9.791
Adam	-0.533	0.999	0.325
Newton-Raphson	-0.504	1.055	0.087
1D Sampling	-0.533	1.070	0.308
MCEM	0.522	1.024	0.309
glmer (lme4)	**-0.519**	**1.019**	**0.307**

Table A.10: Numerical performances for the *seeds* example with MCMC sample size **20** and initial value $(\mathbf{0}, \mathbf{0}, \mathbf{1})$.

	Exec.(s)	Conv.(s)	Conv.(iter.)	loglik diff.	MSE
Fixed 0.05	2.493	NA	NA	236.93937	583.28593
Fixed 0.005	2.491	NA	NA	54.94134	46.09180
Adam	2.671	NA	NA	0.03844	0.00032
Newton-Raphson	4.077	1.849	133	1.85747	0.01671
1D Sampling	18.193	NA	NA	0.03607	0.00092
MCEM	284.575	NA	NA	0.00036	0.00001
glmer (lme4)	0.100	0.100	NA	0.00000	0.00000

Table A.11: Numerical estimates for the *seeds* example with MCMC sample size **20** and initial value $(-\mathbf{1}, -\mathbf{1}, \mathbf{4})$.

	β_0	β_1	σ_{RE}
Fixed 0.05	-0.388	36.539	10.000
Fixed 0.005	-0.532	1.004	0.315
Adam	-0.527	1.033	0.242
Newton-Raphson	-0.516	1.042	0.196
1D Sampling	-0.529	1.050	0.320
MCEM	-0.522	1.024	0.309
glmer (lme4)	**-0.519**	**1.019**	**0.307**

Table A.12: Numerical performances for the *seeds* example with MCMC sample size **20** and initial value $(-\mathbf{1}, -\mathbf{1}, \mathbf{4})$.

	Exec.(s)	Conv.(s)	Conv.(iter.)	loglik diff.	MSE
Fixed 0.05	2.554	NA	NA	117.43856	451.86852
Fixed 0.005	2.332	NA	NA	0.02311	0.00016
Adam	2.522	0.934	117	0.18121	0.00152
Newton-Raphson	4.052	2.226	169	0.51604	0.00429
1D Sampling	18.438	14.198	231	0.01940	0.00040
MCEM	284.575	NA	NA	0.00036	0.00001
glmer (lme4)	0.100	0.100	NA	0.00000	0.00000

Table A.13: Numerical estimates for the *seeds* example with MCMC sample size **300** and initial value $(\mathbf{0}, \mathbf{0}, \mathbf{1})$.

	β_0	β_1	σ_{RE}
Fixed 0.05	-0.663	0.974	1.003
Fixed 0.005	-0.520	1.017	0.329
Adam	-0.533	1.039	0.334
Newton-Raphson	-0.509	1.007	0.316
1D Sampling	-0.521	1.027	0.321
MCEM	-0.522	1.024	0.309
glmer (lme4)	**-0.519**	**1.019**	**0.307**

Table A.14: Numerical performances for the *seeds* example with MCMC sample size **300** and initial value $(\mathbf{0}, \mathbf{0}, \mathbf{1})$.

	Exec.(s)	Conv.(s)	Conv.(iter.)	loglik diff.	MSE
Fixed 0.05	13.321	3.115	69	7.16310	0.16905
Fixed 0.005	13.024	8.251	186	0.01798	0.00016
Adam	12.545	2.626	61	0.03111	0.00044
Newton-Raphson	23.399	NA	NA	0.00516	0.00010
1D Sampling	29.612	11.268	112	0.00820	0.00008
MCEM	284.575	NA	NA	0.00036	0.00001
glmer (lme4)	0.100	0.100	NA	0.00000	0.00000

Table A.15: Numerical estimates for the *seeds* example with MCMC sample size **300** and initial value $(-\mathbf{1}, -\mathbf{1}, \mathbf{4})$.

	β_0	β_1	σ_{RE}
Fixed 0.05	-0.389	1.111	1.939
Fixed 0.005	-0.525	1.018	0.316
Adam	-0.513	1.026	0.300
Newton-Raphson	-0.999	-0.994	10.000
1D Sampling	-0.553	1.030	0.339
MCEM	-0.522	1.024	0.309
glmer (lme4)	**-0.519**	**1.019**	**0.307**

Table A.16: Numerical performances for the *seeds* example with MCMC sample size **300** and initial value $(\mathbf{-1}, \mathbf{-1}, \mathbf{4})$.

	Exec.(s)	Conv.(s)	Conv.(iter.)	loglik diff.	MSE
Fixed 0.05	12.439	NA	NA	16.94627	0.89580
Fixed 0.005	12.439	11.170	268	0.00435	0.00004
Adam	12.236	6.892	168	0.00577	0.00004
Newton-Raphson	22.970	NA	NA	48.06622	32.74360
1D Sampling	30.444	20.947	203	0.06495	0.00077
MCEM	284.575	NA	NA	0.00036	0.00001
glmer (lme4)	0.100	0.100	NA	0.00000	0.00000

A.3 Numerical results for salamander

The *lme4* estimates (in bold) should be viewed as the benchmark for the MLE estimate. Fixed stepsizes, Adam, Newton-Raphson, and 1D sampling did not pass the convergence criterion within 300 iterations. We stopped MCEM at iteration 60 due to the long computational time with little variability between iterations.

Table A.17: Numerical estimates for the *salamander* model with MCMC sample size **300** and initial value $(\mathbf{2}, \mathbf{2}, \mathbf{2}, \mathbf{2}, \mathbf{2}, \mathbf{2})$.

	β_1	β_2	β_3	β_4	σ_F^2	σ_M^2
Fixed 0.05	1.013	0.299	-1.948	0.992	1.388	1.254
Fixed 0.005	1.078	0.345	-1.972	1.070	1.605	1.419
Adam	1.071	0.363	-2.006	1.007	1.468	1.361
Newton-Raphson	1.007	0.302	-1.914	0.949	1.273	1.144
1D Sampling	0.993	0.294	-1.935	1.000	1.430	1.160
MCEM	1.016	0.316	-1.938	1.000	1.377	1.251
glmm	1.023	0.335	-1.908	1.006	1.326	1.221
glmer (lme4)	**1.008**	**0.306**	**-1.896**	**0.990**	**1.174**	**1.041**

Table A.18: Numerical performances for the *salamander* model with MCMC sample size **300** and initial value $(\mathbf{2}, \mathbf{2}, \mathbf{2}, \mathbf{2}, \mathbf{2}, \mathbf{2})$.

	Exec.(s)	loglik diff.	MSE
Fixed 0.05	287.321	0.12660	0.00361
Fixed 0.005	281.251	0.37247	0.01355
Adam	314.740	0.23599	0.00954
Newton-Raphson	577.779	0.04573	0.00107
1D Sampling	387.433	0.11607	0.00297
MCEM	7923.793	0.12066	0.00329
glmm	1181.430	0.08856	0.00220
glmer (lme4)	0.100	0.00000	0.00000

Table A.19: Numerical estimates for the *salamander* model with MCMC sample size **300** and initial value $(\mathbf{4}, \mathbf{4}, \mathbf{4}, \mathbf{4}, \mathbf{4}, \mathbf{4})$.

	β_1	β_2	β_3	β_4	σ_F^2	σ_M^2
Fixed 0.05	1.019	0.323	-1.921	1.058	1.439	1.328
Fixed 0.005	1.757	0.834	-2.027	1.707	3.778	3.518
Adam	0.982	0.338	-1.987	1.049	1.435	1.270
Newton-Raphson	1.003	0.316	-1.930	0.979	1.343	1.184
1D Sampling	1.029	0.306	-2.002	1.029	1.503	1.414
MCEM	1.016	0.316	-1.938	1.000	1.377	1.251
glmm	1.023	0.335	-1.908	1.006	1.326	1.221
glmer (lme4)	**1.008**	**0.306**	**-1.896**	**0.990**	**1.174**	**1.041**

Table A.20: Numerical performances for the *salamander* model with MCMC sample size **300** and initial value $(\mathbf{4}, \mathbf{4}, \mathbf{4}, \mathbf{4}, \mathbf{4}, \mathbf{4})$.

	Exec.(s)	loglik diff.	MSE
Fixed 0.05	286.661	0.20898	0.00608
Fixed 0.005	279.704	6.59919	0.47355
Adam	288.458	0.16740	0.00631
Newton-Raphson	548.292	0.07862	0.00194
1D Sampling	394.195	0.29475	0.01032
MCEM	7923.793	0.12066	0.00329
glmm	1181.430	0.08856	0.00220
glmer (lme4)	0.100	0.00000	0.00000

Appendix B

Supplementary materials for Chapter 3

B.1 Model setup

We propose the following non-parametric model for microbiome data:

$$\text{latent composition:} \quad \boldsymbol{\pi}_i^{(g)}|g \sim f_{\boldsymbol{\pi}}(\cdot|g) \tag{B.1}$$

$$\text{library size:} \quad L_i^{(g)}|\boldsymbol{\pi}_i^{(g)}, g \sim f_L(\cdot|\boldsymbol{\pi}_i^{(g)}, g) \tag{B.2}$$

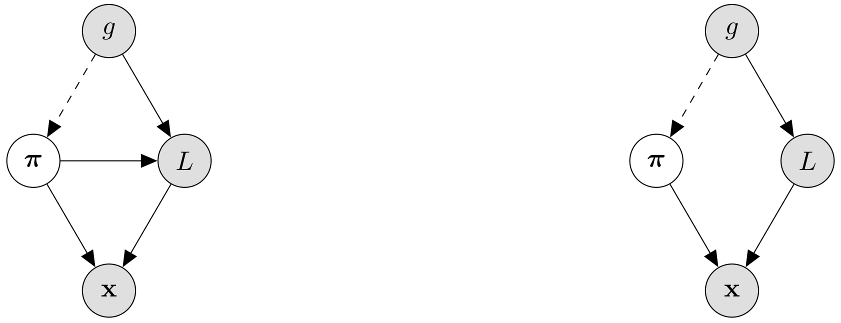
$$\text{count:} \quad \mathbf{x}_i^{(g)}|\boldsymbol{\pi}_i^{(g)}, L_i^{(g)} \sim \text{Multinomial}(L_i^{(g)}, \boldsymbol{\pi}_i^{(g)}), \tag{B.3}$$

where $f_{\boldsymbol{\pi}}(\cdot|g)$ is a probability density function supported on $[0,1]^p$ and $f_L(\cdot|\boldsymbol{\pi}_i^{(g)}, g)$ is a probability mass function supported on non-negative integers. Intuitively, $\boldsymbol{\pi}_i$ represents the latent composition of the sample up to the point of counting sequences. Thus $f_{\boldsymbol{\pi}}(\cdot|g)$ could include effects of extraction, polymerase chain reaction (PCR), and any additional processing procedures. Then the multinomial assumption is that taxa are drawn independently in the machine that does the counting.

B.2 Rarefaction preserves multinomial distribution

The typical rarefaction procedure is as follows:

1. Specify a desired library size L^*.
2. Discard all the samples with library size L_i less than L^*.
3. Subsample all the samples with library size L_i greater than L^* to L^*. This sampling is done via sampling without replacement.



(a) Full model: L depends on both $\boldsymbol{\pi}$ and g.

(b) Simplified model: L depends on only g.

Figure B.1: A non-parametric graphical model for grouped microbiome data. The shaded nodes are observed quantities. The dashed arrow represents the hypothesis of interest: whether microbial compositions vary across different groups.

Since the subsampling is done via sampling without replacement, conditional on the observed count vector $\mathbf{x}$ (and the rarefied depth L^*), the rarefied count vector $\mathbf{x}^*$ follows a multivariate hypergeometric distribution:

$$p(\mathbf{x}^*|\mathbf{x}, L^*) = \frac{\prod_{j=1}^{p} \binom{x_j}{x_j^*}}{\binom{L}{L^*}}, \tag{B.4}$$

where x_j is the jth observed OTU count and x_j^* is the jth rarefied OTU count.

We claim that rarefaction reduces all the library sizes L to a pre-specified rarefying depth L^*, regardless of the group memberships and the latent compositions, while preserving the multinomial distribution of the count vectors. This statement is made rigorous in Theorem 2.

Theorem 2. *Assume the rarefied depth L^* is smaller than or equal to the smallest library size, so no samples are discarded. Let $(\mathbf{x}_1^{(1)*}, ..., \mathbf{x}_{n_1}^{(1)*}, \mathbf{x}_1^{(2)*}, ..., \mathbf{x}_{n_2}^{(2)*})$ be the rarefied data after rarefying the count data $(\mathbf{x}_1^{(1)}, ..., \mathbf{x}_{n_1}^{(1)}, \mathbf{x}_1^{(2)}, ..., \mathbf{x}_{n_2}^{(2)})$ to depth L^*. Under the non-parametric graphical model (B.1) - (B.3), the distribution of $(\mathbf{x}_1^{(1)*}, ..., \mathbf{x}_{n_1}^{(1)*}, \mathbf{x}_1^{(2)*}, ..., \mathbf{x}_{n_2}^{(2)*})$ is described by the following:*

$$\boldsymbol{\pi}_i^{(g)}|g \sim f_{\boldsymbol{\pi}}(\cdot|g) \tag{B.5}$$

$$\mathbf{x}_i^{(g)*}|L^*, \boldsymbol{\pi}_i^{(g)} \sim \text{Multinomial}(L^*, \boldsymbol{\pi}_i^{(g)}). \tag{B.6}$$

Proof. Let $\mathbf{m}_1, \mathbf{m}_2, ...$ be a sequence of iid Multinomial$(1, \boldsymbol{\pi})$ random variable. For all R, $\mathbf{x}_{1:R} := \sum_{t=1}^{R} \mathbf{m}_t$ is Multinomial$(R, \boldsymbol{\pi})$. The original count vector $\mathbf{x}$ conditional on L and $\boldsymbol{\pi}$ can be viewed as $\mathbf{x}_{1:L}$. It remains to show that $\mathbf{x}_{1:L^*}$ conditional on L^* and $\boldsymbol{\pi}$ can be viewed as the rarefied count vector $\mathbf{x}^*$. To this end, we argue that $\mathbf{x}_{1:L^*}|\mathbf{x}_{1:L}, L^*$ follows the multivariate hypergeometric distribution in (B.4), because this implies that $\mathbf{x}_{1:L^*}$ can be generated via sampling $\mathbf{x}$ without replacement, which is exactly rarefaction. We

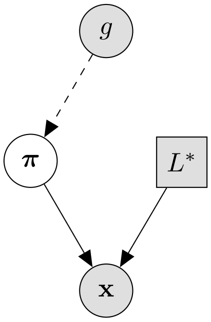


Figure B.2: The graphical model of rarefied data with rarefied depth L^*. A boxed node is used to emphasize that L^* is a fixed value.

denote the jth element of $\mathbf{x}_{1:L}$ by $x_{1:L,j}$ (similar for $\mathbf{x}_{1:L^*}$). For any positive integer K and $\mathbf{v} \in \{\mathbf{w} \in \{0, 1, ..., K\}^p | \sum_{j=1}^p w_j = K\}$, let $\binom{K}{\mathbf{v}} = K!(\prod_{j=1}^p v_j!)^{-1}$ denote the multinomial coefficient. Noting that $L = \sum_{j=1}^p x_{1:L,j}$,

$$\begin{aligned} p(\mathbf{x}_{1:L^*}|\mathbf{x}_{1:L}, L^*, \boldsymbol{\pi}) &= p(\mathbf{x}_{1:L^*}|\mathbf{x}_{1:L}, L^*, \boldsymbol{\pi}, L) \\ &= \frac{\binom{L^*}{\mathbf{x}_{1:L^*}} \prod_{j=1}^p \boldsymbol{\pi}_j^{x_{1:L^*,j}} \binom{L-L^*}{\mathbf{x}_{1:L}-\mathbf{x}_{1:L^*}} \prod_{j=1}^p \boldsymbol{\pi}_j^{x_{1:L,j}-x_{1:L^*,j}}}{\binom{L}{\mathbf{x}_{1:L}} \prod_{j=1}^p \boldsymbol{\pi}_j^{x_{1:L,j}}} \\ &= \frac{\prod_{j=1}^p \binom{x_{1:L,j}}{x_{1:L^*,j}}}{\binom{L}{L^*}}, \end{aligned} \tag{B.7}$$

which does not depend on latent composition $\boldsymbol{\pi}$. Hence, if we marginalize over $\boldsymbol{\pi}$, we get

$$\begin{aligned} p(\mathbf{x}_{1:L^*}|\mathbf{x}, L^*) &= \int p(\mathbf{x}^*|\mathbf{x}, L^*, \boldsymbol{\pi}) p(\boldsymbol{\pi}|\mathbf{x}, L^*) d\boldsymbol{\pi} = \frac{\prod_{j=1}^p \binom{x_{1:L,j}}{x_{1:L^*,j}}}{\binom{L}{L^*}} \int p(\boldsymbol{\pi}|\mathbf{x}, L^*) d\boldsymbol{\pi} \\ &= \frac{\prod_{j=1}^p \binom{x_{1:L,j}}{x_{1:L^*,j}}}{\binom{L}{L^*}}, \end{aligned} \tag{B.8}$$

which is the multivariate hypergeometric distribution in (B.4). □

B.3 How much does rarefaction hurt statistical inference?

Rarefying has been criticized for wasting data since we effectively remove a portion of the data in the downsampling procedure [64]. The subsampling in rarefying will inevitably increase the variance of test statistics and decrease the power of subsequent testing procedures. The key question is how much rarefying hurts statistical inference. We address this question through theoretical considerations and simulation studies. We focus our discussion on two relatively simple testing procedures, the negative binomial Wald test and the two-sample z test. More involved statistical procedures such as DESeq2 and PERMANOVA are hard to analyze theoretically because they integrate various statistical principles. Negative binomial-based tests are used in two popular microbiome analysis routines, edgeR [79] and DESeq2 [59], while the z-test (and the closely related t-test) is a classical statistical test widely used in microbiome studies [100, 104], and it forms an integral part of more involved approaches in the popular package *limma* [77, 71] for differential expression analysis.

B.3.1 Theoretical Considerations

B.3.1.1 Sample relative abundance for an observation

We first consider a single observation, with x being the count for a particular OTU, π being the latent relative abundance of the OTU, and L being the library size. We are interested in the basic statistical properties of the sample OTU relative abundance x/L. Under the multinomial model, $x|\pi, L$ follows a binomial distribution with parameters L and π, with mean $L\pi$ and variance $L\pi(1-\pi)$.

Proposition 1. *The sample relative abundance is an unbiased estimator of the expectation of the latent relative abundance:* $\mathbb{E}\left(\frac{x}{L}\right) = \mathbb{E}(\pi)$.

Proof. Using the law of iterated expectations,

$$\mathbb{E}\left(\frac{x}{L}\right) = \mathbb{E}\left[\mathbb{E}\left[\frac{x}{L}|\pi, L\right]\right] = \mathbb{E}\left[\frac{1}{L}\mathbb{E}\left[x|\pi, L\right]\right] = \mathbb{E}\left[\frac{1}{L}(L\pi)\right] = \mathbb{E}(\pi). \tag{B.9}$$

□

Theorem 3. *The variance of sample relative abundance can be written as the sum of latent decomposition variation and measurement error:*

$$\mathrm{Var}\left(\frac{x}{L}\right) = \underbrace{\mathrm{Var}(\pi)}_{\text{latent variation}} + \underbrace{\mathbb{E}\left[\frac{\pi(1-\pi)}{L}\right]}_{\text{measurement error}}. \tag{B.10}$$

Proof. Using the law of total variance,

$$\mathrm{Var}\left(\frac{x}{L}\right) = \mathrm{Var}\left[\mathbb{E}\left[\frac{x}{L}|\pi, L\right]\right] + \mathbb{E}\left[\mathrm{Var}\left[\frac{x}{L}|\pi, L\right]\right] = \mathrm{Var}(\pi) + \mathbb{E}\left[\frac{\pi(1-\pi)}{L}\right]. \tag{B.11}$$

□

To simplify the rest of our discussion, we will assume the distribution of library sizes does not depend on the latent composition; that is $f_L(\cdot|\boldsymbol{\pi}, g) = f_L(\cdot|g)$, resulting in the graphical model in Figure B.1b. From the graphical model, we can immediately deduce the conditional independence of L and $\boldsymbol{\pi}$, leading to

$$\mathrm{Var}\left(\frac{x}{L}\right) = \underbrace{\mathrm{Var}(\pi)}_{\text{latent variation}} + \underbrace{\mathbb{E}[\pi(1-\pi)]\mathbb{E}\left[\frac{1}{L}\right]}_{\text{measurement error}}. \tag{B.12}$$

The first term in $\mathrm{Var}\left(\frac{x}{L}\right)$ is the variance of the latent relative abundance π, representing the *latent variation*. Samples from the same group do not necessarily have the same OTU proportion due to individual differences. The second term in $\mathrm{Var}\left(\frac{x}{L}\right)$ is the *measurement error*, arise from the sequencing procedure of each sample. The measurement error is proportional to the expected reciprocal of library size $\mathbb{E}(1/L)$. Intuitively if the library size is larger, the precision increases and hence the measurement error decreases.

B.3.1.2 Estimation of expected OTU relative abundance

Next we suppose that there are n observations belonging to the same group. Consider two estimators, the negative binomial estimator and the sample average estimator, for the expected OTU relative abundance p for a particular OTU.

The negative binomial estimator Assume the raw counts follow a negative binomial distribution:

$$x_i \overset{ind}{\sim} NB(\mu_i = L_i p, \phi), \quad i = 1, ..., n. \tag{B.13}$$

The negative binomial estimator is the maximum likelihood estimator of p under the NB model. Using a standard result from large-sample theory, the MLE $\hat{p}_{ML}$ is asymptotically normally distributed with mean p and variance (scaled by $1/n$) approximately the harmonic mean of the conditional variances of relative abundances given the library sizes. See Appendix B.5 for a derivation of this result.

The sample average estimator Without assuming any parametric models, a straightforward estimator of p is the sample average estimator, the arithmetic average of observed proportions:

$$\hat{p}_{\text{avg}} = \frac{1}{n}\sum_{i=1}^{n} \frac{x_i}{L_i}. \tag{B.14}$$

From Proposition 1, each of the observed proportions x_i/L_i is an unbiased estimator of p, so $\hat{p}_{\text{avg}}$ is also an unbiased estimator of p. If we view (x_i, L_i) as independently identically distributed, then by central limit theorem $\hat{p}_{\text{avg}}$ is approximately normal with mean 0 and variance $\mathrm{Var}(x_1/L_1)/n$.

B.3.1.3 Hypothesis testing

Finally we turn to the problem of hypothesis testing. Let $n = n_1 + n_2$ be the total sample size. Assume that $\lim_{n\to\infty} n_1/n = c \in (0,1)$. Suppose we have two groups of independent samples $((x_1^{(1)}, L_1^{(1)}), ..., (x_{n_1}^{(1)}, L_{n_1}^{(1)}))$ and $((x_1^{(2)}, L_1^{(2)}), ..., (x_{n_2}^{(2)}, L_{n_2}^{(2)}))$. Assume that $\mathbb{E}(x_1^{(1)}/L_1^{(1)}) = ... = \mathbb{E}(x_{n_1}^{(1)}/L_{n_1}^{(1)}) = p^{(1)}$ and $\mathbb{E}(x_1^{(2)}/L_1^{(2)}) = ... = \mathbb{E}(x_{n_2}^{(2)}/L_{n_2}^{(2)}) = p^{(2)}$. The null hypothesis of interest is $H_0 : p^{(1)} = p^{(2)}$.

We would like to compare the theoretical performance of a testing procedure based on original data versus the same procedure based on rarefied data. To do so, we make use of the notion of asymptotic relative efficiency (ARE) [89]. Intuitively, ARE measures how many observations are needed for the first test compared to the second test, while holding fixed the desired significance level and the lower bound of power.

Definition 1. *Let $w_k(\theta; n)$ be the power function, based on n observations, for test k, $k = 1, 2$, of $H_0 : \theta = 0$ against the alternative $H_1 : \theta = \theta_\nu$, where the sequence of alternatives is indexed by ν with $\theta_\nu \to 0$ as $\nu \to \infty$. For $k = 1, 2$, let $n_{\nu,k}$ be the minimal number of observations such that $w_k(0; n_{\nu,k}) \leq \alpha \in (0, 1)$ and $w_k(\theta_\nu; n_{\nu,k}) \geq \gamma \in (\alpha, 1)$. Suppose the limit*

$$ARE = \lim_{\nu\to\infty} \frac{n_{\nu,2}}{n_{\nu,1}} \tag{B.15}$$

exists. Then ARE is called the asymptotic relative efficiency (or Pitman efficiency) of the first with respect to the second sequence of tests.

The following lemma (Theorem 14.19 in [89]) simplifies our studies of ARE.

Lemma 1. *Consider statistical models $(P_{n,\theta} : \theta \geq 0)$ such that $||P_{n,\theta} - P_{n,0}|| \overset{\theta\to 0}{\to} 0$ for every n. For $k = 1, 2$, suppose the sequence of statistics $T_{n,k}$ is asymptotically normal in the sense that for all sequences $\theta_n \downarrow 0$,*

$$\frac{\sqrt{n}(T_{n,k} - \mu_k(\theta_n))}{\sigma_k(\theta_n)} \overset{\theta_n}{\rightsquigarrow} N(0, 1), \tag{B.16}$$

where $\overset{\theta_n}{\rightsquigarrow}$ is the law indexed by θ_n, μ_k is differentiable at zero and σ_k is continuous at zero with $\mu_k'(0) > 0$ and $\sigma_k(0) > 0$. The ARE of the tests that reject the null hypothesis $H_0 : \theta = 0$ for large values of $T_{n,k}$ is equal to

$$\left(\frac{\mu_1'(0)/\sigma_1(0)}{\mu_2'(0)/\sigma_2(0)}\right)^2. \tag{B.17}$$

for every sequence of alternatives $\theta_\nu \downarrow 0$, independently of $\alpha \in (0, 1)$ and $\gamma \in (\alpha, 1)$.

In our setting, tests $k = 1, 2$ refer to the test based on rarefied data and the test based on original data respectively. The test statistic $T_{n,k}$ is the difference in estimated proportions for the two groups, $\theta_n = p_1^{(g)} - p_2^{(g)}$, $\mu_k(\theta_n) = \theta_n$ (so $\mu_k'(0) = 1$), and $\sigma_k^2(\theta_n) = \mathrm{Var}(T_{n,k})$. It can be shown that for the NB Wald test and the two-sample z-test the ARE can be written in terms of ratios of variances $\mathrm{Var}(\frac{x_i}{L_i})/\mathrm{Var}(\frac{x^*}{L^*})$ (see Theorems 5 and 6 in Appendix B.6).

Under the nonparametric graphical model, the ratios of variances $\mathrm{Var}(\frac{x_i}{L_i})/\mathrm{Var}(\frac{x^*}{L^*})$ can be written approximately in terms of the variance-to-mean ratio (VMR), also known as coefficient of dispersion, of the latent relative abundance, as well as the expectation of the reciprocal of library size. We denote the VMR of π as $\mathrm{VMR}(\pi) := \mathrm{Var}(\pi)/\mathbb{E}(\pi)$.

Theorem 4. *Under the non-parametric graphical model (B.1) - (B.3), provided that the distribution of library sizes does not depend on the latent composition, the ratio of variances for original data (x, L) and rarefied data (x^*, L^*) is*

$$\frac{\mathrm{Var}(\frac{x}{L})}{\mathrm{Var}(\frac{x^*}{L^*})} = 1 - \left[\frac{1}{L^*} - \mathbb{E}\left(\frac{1}{L}\right)\right]\left[\frac{\mathrm{Var}(\pi)}{\mathbb{E}[\pi(1-\pi)]} + \frac{1}{L^*}\right]^{-1} \approx 1 - \frac{\frac{1}{L^*} - \mathbb{E}\left(\frac{1}{L}\right)}{VMR(\pi) + \frac{1}{L^*}}, \tag{B.18}$$

where the approximation is reasonable if π is small with high probability.

Proof. The claim follows immediately from Theorem 3. □

To gain intuition from Theorem 4, we consider two scenarios for the VMR of the latent relative abundance:

1. $\mathrm{VMR}(\pi) \ll 1/L^*$: The VMR of the latent relative abundance is much smaller than the reciprocal of rarefied depth. In this case, the ratio of variances is approximately

$$\left[\mathbb{E}\left(\frac{1}{L}\right)\right] / \left(\frac{1}{L^*}\right), \tag{B.19}$$

   so by Theorems 5 and 6 in Appendix B.6 the AREs of both the NB Wald test and the two-sample z-test are purely governed by the measurement error. In this case rarefying can potentially weaken the power of the testing procedure substantially.

2. $\mathrm{VMR}(\pi) \gg \frac{1}{L^*} - \mathbb{E}\left(\frac{1}{L}\right)$: The VMR of the latent relative abundance is much greater than loss in precision due to rarefying. In this case the ratio of variances is close to 1, so measurement error does not matter and reasonable rarefying procedures would not affect the ARE much.

B.4 Dirichlet-multinomial model and the negative binomial distribution

In this section, we show that a Dirichlet-multinomial simulation will approximately satisfy negative binomial analysis assumptions.

Suppose

$$\boldsymbol{\pi} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}), \quad \mathbf{x}|L, \boldsymbol{\pi} \sim \mathrm{Multinomial}(L, \boldsymbol{\pi}),$$

where $\boldsymbol{\alpha} \in \mathbb{R}^J$ and $L > 0$ is the library size. Consider a particular taxon, say j. From the Dirichlet-multinomial model,

$$\pi_j \sim \text{Beta}(\alpha_j, \sum_{k \neq j} \alpha_k), \quad x_j | L, \pi_j \sim \text{Binomial}(L, \pi_j).$$

In the context of microbiome studies π_j tends to be small, so $x_j|L, \pi_j$ is approximately Poisson distributed. On the other hand, using $1 - y \approx e^{-y}$ for small y, the probability density function of π_j is

$$f(\pi_j) \propto \pi_j^{\alpha_j - 1}(1 - \pi_j)^{\sum_{k \neq j} \alpha_k - 1} \approx \pi_j^{\alpha_j - 1} e^{-(\sum_{k \neq j} \alpha_k - 1)\pi_j},$$

so π_j is approximately Gamma distributed. Since π_j is approximately Gamma and $x_j|L, \pi_j$ is approximately Poisson, the marginal distribution of x_j is approximately negative binomial.

B.5 Asymptotic distribution of the negative binomial estimator

Suppose $x_1, ..., x_n$ are independent $NB(\mu_i = L_i p, \phi)$, with L_i being fixed. The corresponding log-likelihood is

$$l(p, \phi) = \sum_{j=1}^{n} \left[\log \binom{x_i + \phi - 1}{x_i} + x_i \log \left(\frac{L_i p}{L_i p + \phi} \right) + \phi \log \left(\frac{\phi}{L_i p + \phi} \right) \right]. \tag{B.20}$$

Let $(\hat{p}_{ML}, \hat{\phi}_{ML})$ be the maximum likelihood estimate, obtained by maximizing $l(p, \phi)$. While there is no closed form solution for $(\hat{p}_{ML}, \hat{\phi}_{ML})$, we can gain some intuition of $\hat{p}_{ML}$ from the partial derivatives of $l(p, \phi)$. Note that

$$\frac{\partial}{\partial p} l(p, \phi) = \phi \sum_{j=1}^{n} \frac{x_i - L_i p}{p(L_i p + \phi)} = \sum_{j=1}^{n} L_i^2 \frac{\frac{x_i}{L_i} - p}{\left(L_i p + \frac{L_i^2 p^2}{\phi} \right)} = \sum_{j=1}^{n} \frac{\frac{x_i}{L_i} - p}{\text{Var}(x_i)/L_i^2} \tag{B.21}$$

$$= \sum_{j=1}^{n} \frac{1}{\text{Var}(x_i/L_i)} \left(\frac{x_i}{L_i} - p \right) \tag{B.22}$$

From the first order condition $\frac{\partial}{\partial p} l(p, \phi) \big|_{(\hat{p}_{ML}, \hat{\phi}_{ML})} = 0$, we get

$$\hat{p}_{ML} = \frac{1}{\sum_{j=1}^{n} \frac{1}{\widehat{\text{Var}}(x_i/L_i)}} \sum_{j=1}^{n} \frac{1}{\widehat{\text{Var}}(x_i/L_i)} \left(\frac{x_i}{L_i} \right) \approx \frac{1}{\sum_{j=1}^{n} \frac{1}{\text{Var}(x_i/L_i)}} \sum_{j=1}^{n} \frac{1}{\text{Var}(x_i/L_i)} \left(\frac{x_i}{L_i} \right), \tag{B.23}$$

where $\widehat{\text{Var}}(x_i/L_i) = L_i \hat{p}_{ML} + \frac{L_i^2 \hat{p}_{ML}^2}{\hat{\phi}_{ML}}$, so $\hat{p}_{ML}$ is approximately an inverse-variance-weighted average of observed proportions.

On the other hand, the second partial derivatives of $l(p, \phi)$ with respect to p is

$$\frac{\partial^2}{\partial p^2} l(p, \phi) = \phi \sum_{j=1}^{n} \frac{L_i^2 p^2 - x_i(2L_i p + \phi)}{p^2 (L_i p + \phi)^2}. \tag{B.24}$$

Therefore,

$$-E\left[\frac{\partial^2}{\partial p^2} l(p, \phi)\right] = \phi \sum_{j=1}^{n} \frac{L_i p(2L_i p + \phi) - L_i^2 p^2}{p^2 (L_i p + \phi)^2} = \phi^2 \sum_{j=1}^{n} \frac{L_i p + \frac{L_i^2 p^2}{\phi}}{p^2 (L_i p + \phi)^2} \tag{B.25}$$

$$= \sum_{j=1}^{n} \frac{\mathrm{Var}(x_i)}{p^2 \left(\frac{L_i p}{\phi} + 1\right)^2} = \sum_{j=1}^{n} L_i^2 \frac{\mathrm{Var}(x_i)}{\left(\frac{L_i^2 p^2}{\phi} + L_i p\right)^2} = \sum_{j=1}^{n} \frac{L_i^2}{\mathrm{Var}(x_i)} \tag{B.26}$$

$$= \sum_{j=1}^{n} \frac{1}{\mathrm{Var}(x_i/L_i)}. \tag{B.27}$$

Using a standard result from large-sample theory, the MLE $\hat{p}_{ML}$ is asymptotically normally distributed with mean p and variance

$$\left(-E\left[\frac{\partial^2}{\partial p^2} l(p, \phi)\right]\right)^{-1} = \frac{1}{\sum_{j=1}^{n} \frac{1}{\mathrm{Var}(x_i/L_i)}} = \frac{1}{n} HM\left(\mathrm{Var}\left(\frac{x_1}{L_1}\right), ..., \mathrm{Var}\left(\frac{x_n}{L_n}\right)\right), \tag{B.28}$$

where $HM(a_1, ..., a_n) = \frac{n}{\frac{1}{a_1} + ... + \frac{1}{a_n}}$ is the harmonic mean of $a_1, ..., a_n$.

Now we turn to the case that L_i is viewed as random, with a distribution that does not depend on p or ϕ. The MLE is still $\hat{p}_{ML}$, but the asymptotic variance is different. Using law of iterated expectations and the modeling assumption (x_i, L_i) being independently identically distributed,

$$\begin{aligned} -\mathbb{E}\left[\frac{\partial^2}{\partial p^2} l(p, \phi)\right] &= -\mathbb{E}\left[\mathbb{E}\left[\frac{\partial^2}{\partial p^2} l(p, \phi)\middle| L_1, ..., L_n\right]\right] = \mathbb{E}\left[\sum_{i=1}^{n} \frac{1}{\mathrm{Var}(x_i/L_i|L_i)}\right] \\ &= n\mathbb{E}\left[\frac{1}{\mathrm{Var}(x_i/L_i|L_i)}\right]. \end{aligned} \tag{B.29}$$

Using law of total variances,

$$\begin{aligned} \mathrm{Var}\left(\frac{x_i}{L_i}\right) &= \mathbb{E}\left[\mathrm{Var}\left(\frac{x_i}{L_i}\middle| L_i\right)\right] + \mathrm{Var}\left[\mathbb{E}\left(\frac{x_i}{L_i}\middle| L_i\right)\right] = \mathbb{E}\left[\mathrm{Var}\left(\frac{x_i}{L_i}\middle| L_i\right)\right] + \mathrm{Var}(p) \\ &= \mathbb{E}\left[\mathrm{Var}\left(\frac{x_i}{L_i}\middle| L_i\right)\right]. \end{aligned} \tag{B.30}$$

Putting all the pieces together, the asymptotic variance is

$$\left(-\mathbb{E}\left[\frac{\partial^2}{\partial p^2} l(p, \phi)\right]\right)^{-1} = \frac{1}{n} \frac{1}{\mathbb{E}\left[\frac{1}{\mathrm{Var}(x_i/L_i|L_i)}\right]} \leq \frac{1}{n} \frac{1}{\frac{1}{\mathbb{E}\mathrm{Var}(x_i/L_i|L_i)}} = \frac{\mathrm{Var}(x_i/L_i)}{n}, \tag{B.31}$$

where the inequality is due to Jensen's inequality.

B.6 Hypothesis testing

In this appendix, we show that the AREs for both the negative binomial Wald test and the two-sample z-test can be expressed in terms of ratios of variances $\mathrm{Var}(x/L)/\mathrm{Var}(x^*/L^*)$.

The negative binomial Wald test Let $\hat{p}_{ML}^{(1)}$ and $\hat{p}_{ML}^{(2)}$ be the NB estimators of proportions computed based on the first group of samples and the second group respectively. We define the negative binomial test of $H_0 : p^{(1)} = p^{(2)}$ to be the test that rejects H_0 if $|\hat{p}_{ML}^{(1)} - \hat{p}_{ML}^{(2)}|/\widehat{\mathrm{SE}}(\hat{p}_{ML}^{(1)} - \hat{p}_{ML}^{(2)}) > z_{\alpha/2}$, where $\widehat{\mathrm{SE}}(\hat{p}_{ML}^{(1)} - \hat{p}_{ML}^{(2)})$ is the estimated standard error of $\hat{p}_{ML}^{(1)} - \hat{p}_{ML}^{(2)}$, α is the desired significance level, and $z_{\alpha/2}$ is the upper $\alpha/2$-quantile of a standard normal distribution.

We first state the result for the case of fixed library sizes and then state a corollary for the case of random library sizes.

Theorem 5. *Suppose $x_i^{(g)}$ are independent $NB(L_i^{(g)}p^{(g)}, \phi)$ for $i = 1, ..., n_g$, $g = 1, 2$ with the library sizes $L_i^{(g)}$ being fixed values. Suppose $c := \lim_{n\to\infty} n_1/n \in (0, 1)$, and for $k = 1, 2$ the limit*

$$\lim_{n\to\infty} HM\left(\mathrm{Var}\left(\frac{x_1^{(g)}}{L_1^{(g)}}\right), ..., \mathrm{Var}\left(\frac{x_{n_g}^{(g)}}{L_{n_g}^{(g)}}\right)\right) \tag{B.32}$$

exists, where $HM(a_1, ..., a_{n_g}) = \frac{n_g}{\frac{1}{a_1}+...+\frac{1}{a_{n_g}}}$ is the harmonic mean of $(a_1, ..., a_{n_g})$. Suppose the rarefied depth $L^* \le L_i^{(k)}$ for all j, k, so no samples are discarded. Assume that in the limit of any sequence of alternatives considered $p^{(g)} \to p$ for $g = 1, 2$. The asymptotic relative efficiency of the negative binomial test based on the rarefied data with rarefied depth L^* relative to the same test based on the original data is

$$\left[(1-c)\lim_{n\to\infty} HM\left(\frac{\mathrm{Var}(x_1^{(1)'}/L_1^{(1)})}{\mathrm{Var}(x^*/L^*)}, ..., \frac{\mathrm{Var}(x_{n_1}^{(1)'}/L_{n_1}^{(1)})}{\mathrm{Var}(x^*/L^*)}\right)\right. \tag{B.33}$$

$$\left. + c\lim_{n\to\infty} HM\left(\frac{\mathrm{Var}(x_1^{(2)'}/L_1^{(2)})}{\mathrm{Var}(x^*/L^*)}, ..., \frac{\mathrm{Var}(x_{n_2}^{(2)'}/L_{n_2}^{(2)})}{\mathrm{Var}(x^*/L^*)}\right)\right], \tag{B.34}$$

where $x_i^{(g)'} \sim NB(L_i^{(g)}p, \phi)$ for $g = 1, 2$ and $x^* \sim NB(L^*p, \phi)$.

Corollary 1. *Suppose $(x_i^{(g)}, L_i^{(g)})$ are independent with*

$$L_i^{(g)} \sim f^{(g)} \tag{B.35}$$

$$x_i^{(g)}|L_i^{(g)} \sim NB(L_i^{(g)}p^{(g)}, \phi) \tag{B.36}$$

for $i = 1, ..., n_g$, $g = 1, 2$ with $f^{(g)}$ being some discrete distribution on non-negative integers. Suppose $c := \lim_{n\to\infty} n_1/n \in (0, 1)$. Suppose further the rarefied depth $L^* \le L_i^{(k)}$ for all

j, k, so no samples are discarded. Assume that in the limit of any sequence of alternatives considered $p^{(g)} \to p$ for $g = 1, 2$. The asymptotic relative efficiency of the negative binomial test based on the rarefied data with rarefied depth L^* relative to the same test based on the original data is

$$\frac{1}{Var(x^*/L^*)}\left[(1-c)\frac{1}{\mathbb{E}\left(\frac{1}{Var(x_i^{(1)'}/L_i^{(1)}|L_i^{(1)})}\right)}+c\frac{1}{\mathbb{E}\left(\frac{1}{Var(x_i^{(2)'}/L_i^{(2)}|L_i^{(2)})}\right)}\right], \tag{B.37}$$

where $x_i^{(g)'}|L_i^{(g)} \sim NB(L_i^{(g)}p, \phi)$ for $g = 1, 2$ and $x^* \sim NB(L^*p, \phi)$.

Note that the expression in B.37 is upper bounded by

$$(1-c)\frac{V\mathrm{ar}(x_i^{(1)'}/L_i^{(1)})}{V\mathrm{ar}(x^*/L^*)}+c\frac{V\mathrm{ar}(x_i^{(2)'}/L_i^{(2)})}{V\mathrm{ar}(x^*/L^*)}, \tag{B.38}$$

since

$$\begin{aligned}\mathrm{Var}\left(\frac{x_i^{(1)'}}{L_i^{(1)}}\right) &= \mathbb{E}\left(\mathrm{Var}\left(\frac{x_i^{(1)'}}{L_i^{(1)}}\middle| L_i^{(1)}\right)\right)+\mathrm{Var}\left(\mathbb{E}\left(\frac{x_i^{(1)'}}{L_i^{(1)}}\middle| L_i^{(1)}\right)\right)\\ &= \mathbb{E}\left(\mathrm{Var}\left(\frac{x_i^{(1)'}}{L_i^{(1)}}\middle| L_i^{(1)}\right)\right)+\mathrm{Var}(p)\\ &= \mathbb{E}\left(\mathrm{Var}\left(\frac{x_i^{(1)'}}{L_i^{(1)}}\middle| L_i^{(1)}\right)\right)\end{aligned} \tag{B.39}$$

and by Jensen's inequality

$$\frac{1}{\mathbb{E}\left(\frac{1}{V\mathrm{ar}(x_i^{(g)'}/L_i^{(1)}|L_i^{(g)})}\right)} \le \mathbb{E}\left(\mathrm{Var}\left(\frac{x_i^{(g)'}}{L_i^{(g)}}\middle| L_i^{(g)}\right)\right) = \mathrm{Var}\left(\frac{x_i^{(g)'}}{L_i^{(g)}}\right) \tag{B.40}$$

for $g \in \{1, 2\}$.

The two-sample z-test Let $\hat{p}_{\text{avg}}^{(1)}$ and $\hat{p}_{\text{avg}}^{(2)}$ be the sample average estimators of proportions computed based on the first group of samples and the second group respectively. We define the two-sample z-test of $H_0 : p^{(1)} = p^{(2)}$ to be the test that rejects H_0 if $|\hat{p}_{\text{avg}}^{(1)} - \hat{p}_{\text{avg}}^{(2)}|/\widehat{\mathrm{SE}}(\hat{p}_{\text{avg}}^{(1)} - \hat{p}_{\text{avg}}^{(2)}) > z_{\alpha/2}$, , where $\widehat{\mathrm{SE}}(\hat{p}_{\text{avg}}^{(1)} - \hat{p}_{\text{avg}}^{(2)})$ is the estimated standard error of $\hat{p}_{\text{avg}}^{(1)} - \hat{p}_{\text{avg}}^{(2)}$, α is the desired significance level, and $z_{\alpha/2}$ is the upper $\alpha/2$-quantile of a standard normal distribution.

Similar to the NB Wald test result, we first state the result for the case of fixed library sizes and then state a corollary for the case of random library sizes.

Theorem 6. *Suppose $x_i^{(g)}$ follows a certain distribution $P_{L_i^{(g)},p^{(g)}}$, indexed by the library size and the expected relative abundance. Assume that library sizes are fixed. Further assume that $c := \lim_{n\to\infty} n_1/n \in (0,1)$, and for $k = 1, 2$ the limit*

$$\lim_{n\to\infty} AM\left(\mathrm{Var}\left(\frac{x_1^{(g)}}{L_1^{(g)}}\right), ..., \mathrm{Var}\left(\frac{x_{n_g}^{(g)}}{L_{n_g}^{(g)}}\right)\right) \tag{B.41}$$

exists, where $AM(a_1, ..., a_{n_g}) = (a_1 + ... + a_{n_g})/n_g$ is the arithmetic mean of $(a_1, ..., a_{n_g})$. Suppose the rarefied depth $L^* \le L_i^{(k)}$ for all j, k, so no samples are discarded. Assume that in the limit of any sequence of alternatives considered $p^{(g)} \to p$ and for all $i = 1, ..., n_g$ $||P_{L_i^{(g)},p^{(g)}} - P_{L_i^{(g)},p}|| \to 0$ for $g \in \{1,2\}$. The asymptotic relative efficiency of the z-test based on the rarefied data with rarefied depth L^* relative to the same test based on the original data is

$$\left[(1-c)\lim_{n\to\infty} AM\left(\frac{\mathrm{Var}(x_1^{(1)'}/L_1^{(1)})}{\mathrm{Var}(x^*/L^*)}, ..., \frac{\mathrm{Var}(x_{n_1}^{(1)'}/L_{n_1}^{(1)})}{\mathrm{Var}(x^*/L^*)}\right)\right. \tag{B.42}$$

$$\left. + c\lim_{n\to\infty} AM\left(\frac{\mathrm{Var}(x_1^{(2)'}/L_1^{(2)})}{\mathrm{Var}(x^*/L^*)}, ..., \frac{\mathrm{Var}(x_{n_2}^{(2)'}/L_{n_2}^{(2)})}{\mathrm{Var}(x^*/L^*)}\right)\right], \tag{B.43}$$

where $x_i^{(g)'} \sim P_{L_i^{(g)},p}$ for $g \in \{1,2\}$ and $x^* \sim P_{L^*,p}$.

Corollary 2. *Suppose $x_i^{(g)}|L_i^{(g)}$ follows a certain distribution $P_{L_i^{(g)},p^{(g)}}$, indexed by the library size and the expected relative abundance. Assume that in the limit of any sequence of alternatives considered $p^{(g)} \to p$ and for all $i = 1, ..., n_g$ $||P_{L_i^{(g)},p^{(g)}} - P_{L_i^{(g)},p}|| \to 0$ for $g \in \{1,2\}$. The asymptotic relative efficiency of the z-test based on the rarefied data with rarefied depth L^* relative to the same test based on the original data is*

$$(1-c)\frac{Var(x_i^{(1)'}/L_i^{(1)})}{Var(x^*/L^*)} + c\frac{Var(x_i^{(2)'}/L_i^{(2)})}{Var(x^*/L^*)} \tag{B.44}$$

where $x_i^{(g)'}|L_i^{(g)} \sim P_{L_i^{(g)},p}$ for $g \in \{1,2\}$ and $x^* \sim P_{L^*,p}$.

Remarks

1. The ARE of using rarefied data versus original data for the z-test are completely determined by the ratios of variances $\mathrm{Var}(\frac{x_i^{(g)}}{L_i^{(g)}})/\mathrm{Var}(\frac{x^*}{L^*})$.

2. Suppose the NB model holds. Since AM is at least as large as HM, in terms of ARE, rarefying hurts us more for the NB Wald test than for the two-sample z-test.

B.7 Proof of Theorem 5 and Theorem 6

Now we prove Theorem 5. The proof of Theorem 6 is similar and hence is omitted.

Proof. For $k = 1, 2$, let $\hat{p}^{(k)}_{ML}$ be the NB estimator of the proportion for group k based on the original data and $\hat{p}^{(k)}_{ML,r}$ be the same estimator based on rarefied data. Under the null hypothesis,

$$\frac{\sqrt{n}[(\hat{p}^{(1)}_{ML} - \hat{p}^{(2)}_{ML}) - (p^{(1)} - p^{(2)})]}{\sqrt{H}} \xrightarrow{d} N(0, 1), \tag{B.45}$$

where

$$\begin{aligned} H = &\frac{1}{c} \lim_{n\to\infty} HM \left(\mathrm{Var}\left(\frac{x^{(1)}_1}{L^{(1)}_1}\right), ..., \mathrm{Var}\left(\frac{x^{(1)}_{n_1}}{L^{(1)}_{n_1}}\right)\right) \\ &+ \frac{1}{1-c} \lim_{n\to\infty} HM \left(\mathrm{Var}\left(\frac{x^{(2)}_1}{L^{(2)}_1}\right), ..., \mathrm{Var}\left(\frac{x^{(2)}_{n_2}}{L^{(2)}_{n_2}}\right)\right). \end{aligned} \tag{B.46}$$

On the other hand, under the null hypothesis,

$$\frac{\sqrt{n}[(\hat{p}^{(1)}_{ML,r} - \hat{p}^{(2)}_{ML,r}) - (p^{(1)} - p^{(2)})]}{\sqrt{\left(\frac{1}{c} + \frac{1}{1-c}\right) \mathrm{Var}\left(\frac{x^*}{L^*}\right)}} \xrightarrow{d} N(0, 1). \tag{B.47}$$

Putting everything together, by Lemmma 1, the ARE is

$$\frac{H}{\left(\frac{1}{c} + \frac{1}{1-c}\right) \mathrm{Var}\left(\frac{x^*}{L^*}\right)} = \Bigg[(1-c) \lim_{n\to\infty} HM \left(\frac{\mathrm{Var}(x^{(1)}_1/L^{(1)}_1)}{\mathrm{Var}(x^*/L^*)}, ..., \frac{\mathrm{Var}(x^{(1)}_{n_1}/L^{(1)}_{n_1})}{\mathrm{Var}(x^*/L^*)}\right) \tag{B.48}$$

$$+ c \lim_{n\to\infty} HM \left(\frac{\mathrm{Var}(x^{(2)}_1/L^{(2)}_1)}{\mathrm{Var}(x^*/L^*)}, ..., \frac{\mathrm{Var}(x^{(2)}_{n_2}/L^{(2)}_{n_2})}{\mathrm{Var}(x^*/L^*)}\right)\Bigg]. \tag{B.49}$$

□

B.8 Estimation of variance of rarefied relative abundance

A natural estimator of $\mathrm{Var}(x^{(g)}/L^{(g)})$ is the sample variance of observed relative abundance:

$$S^{(g)}_j := \frac{1}{n_g - 1} \sum_{i=1}^{n_g} \left(\frac{x^{(g)}_{ij}}{L^{(g)}_i} - \frac{1}{n_g} \sum_{l=1}^{n_g} \frac{x^{(g)}_{ij}}{L^{(g)}_l} \right)^2. \tag{B.50}$$

Similarly, a straightforward estimator of $\mathrm{Var}(x^*/L^*)$ is the sample variance of observed relative abundance based on rarefied data. However, the realized value of this estimator $\mathrm{Var}(x^*/L^*)$

can be larger than $S_j^{(g)}$, which is undesirable because $\mathrm{Var}(x^{(g)}/L^{(g)})/\mathrm{Var}(x^*/L^*)$ is always at most 1. In addition, this estimator is a function of the rarefied data, so there is additional variance introduced in this estimator by the subsampling procedure.

We would like to construct an estimator of $\mathrm{Var}(x^*/L^*)$ that depends on the original data. By the law of total variance, conditioning on x and L,

$$\begin{aligned}\mathrm{Var}\left(\frac{x^*}{L^*}\right) &= \mathrm{Var}\left(\mathbb{E}\left(\frac{x^*}{L^*}\middle| x^{(g)}, L^{(g)}\right)\right) + \mathbb{E}\left(\mathrm{Var}\left(\frac{x^*}{L^*}\middle| x^{(g)}, L^{(g)}\right)\right) \\ &= \mathrm{Var}\left(\frac{x^{(g)}}{L^{(g)}}\right) + \mathbb{E}\left(\frac{1}{L^*}\frac{x^{(g)}}{L^{(g)}}\left(1 - \frac{x^{(g)}}{L^{(g)}}\right)\left(\frac{L^{(g)} - L^*}{L^{(g)} - 1}\right)\right).\end{aligned} \tag{B.51}$$

The first term $\mathrm{Var}(x^{(g)}/L^{(g)})$ in (B.51) can be estimated via $S_j^{(g)}$; the second term can be estimated by an empirical average:

$$V_j^{(g)}(L^*) := \frac{1}{n_g L^*}\sum_{i=1}^{n_g} \frac{x_{ij}^{(g)}}{L_i^{(g)}}\left(1 - \frac{x_{ij}^{(g)}}{L_i^{(g)}}\right)\left(\frac{L_i^{(g)} - L^*}{L_i^{(g)} - 1}\right), \tag{B.52}$$

using the fact that the conditional distribution of x^* given (x, L) is a Hypergeometric distribution. Intuitively, the quantity $V_j^{(g)}(L^*)$ estimates the additional variance induced by rarefying the data to depth L^*. Note that $V_j^{(g)}(L^*)$ is a decreasing function in the rarefied depth L^*, which is consistent with the intuition that a larger rarefied depth preserves more data and hence less variability is being introduced by rarefaction. From this discussion, we see that $\mathrm{Var}(x^*/L^*)$ can be estimated by $S_j^{(g)} + V_j^{(g)}(L^*)$, a quantity depending on the original data but not the rarefied data.