Randy

# LDTM

The goal of LDTM is to …

## Installation

You can install the development version of LDTM from [GitHub](https://github.com/Goodgolden/LDTM) with:

# install.packages("devtools")  
devtools::install\_github("Goodgolden/LDTM")

## Example

This is a basic example which shows you how to solve a common problem:

library(LDTM)  
#> Loading required package: tidyverse  
#> ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──  
#> ✓ ggplot2 3.3.5 ✓ purrr 0.3.4  
#> ✓ tibble 3.1.6 ✓ dplyr 1.0.8  
#> ✓ tidyr 1.2.0 ✓ stringr 1.4.0  
#> ✓ readr 2.1.2 ✓ forcats 0.5.1  
#> ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
#> x dplyr::filter() masks stats::filter()  
#> x dplyr::lag() masks stats::lag()  
## basic example code

## Introduction

### OTU operational taxonomic units

microbial genomic sequencess clustered by sequence similarity

partition sequences into discrete groups instead of traditional taxonomic units.

the most abundant sequence in an OTU is the representative sequence

representative sequences from all the OTUs are used to construct a phylogenetic tree among all the OTUs

Microbial community information == OTUs + counts + phylogenetic relationship + taxonomy

### Human gut microbiome study in University of Pennsylvania

the effect of diet on gut microbiome composition

* gut microbiome data
* nutrient intake data

### distance based analysis

* identify a few gut microbiome associated nutrients
* unable to provide information on how dietary nutrient affect bacterial taxa

## Goal

identify both the key nutrients as well as the taxa the nutrient affect

Chen and Li (2013) adopted a regression-based approach

OTU abundance data as multivariate count responses, and nutrient as covaraite

## Model

### Multinomial logistic regression

The link function is a multinomial-Poisson transformation

**might need to use Poissonization to simulate the data**

#### Overdisperson

* the multinomial distribution is not appropriate
* is a random variable with some prior distribution
* is a (d-1) dimensional simplex,

the support of p is

* Dirichlet distribution density at
* Dirichlet is a conjugate prior to multinomial distribution
* posterior is Dirichlet multionmial distribution, aka the Dirichlet compound multinomial distribution.

#### Limitation

* all componnets must share a common variance parameter
* components are mutually independent, up to the constraint that must sum up to 1
* distribution fails to take into account the special and inherent property of microbiome count data (evolutionary relationships in the phylogenetic tree)

#### Generalization of Dirichlet Multinomial distribution

* the relationships among the components of the count vector can be represented as a tree, node-by-node.
  + each component has a independent variance
  + components are correlated at subtree levels
* a regression model with the effects of covariates
* a regularized methods for selecting covarites (nutrients) that are associated with the count responses (OTUs)

#### Logistic Normal Multinomial distribution

* Billheimer with Aitchison’s logistic normal distribution instead of Dirichlet
  + covaraites to the count vector (Billheimer et al. 2001)
  + link dietary nutrients with bacteria counts (Xia et al. 2013)

cannot exploit the tree structure information, logistic normal multinomial does not have a closed-form expression.

### Dirichlet-Tree Multinomial Distributions

total number of counts, determined by the sequence depth, as an ancillary statistics

* analyssi conditioning on this number

A Tree representing the hierarchical structure over the count responses

* we will incorperate the tree structural information into the modeling
* the set of leaf nodes
* the set of internal nodes of
* for each , the child nodes of
* for each and ,
* is the index for the subtree counts and probabilities

* Let to be the root node of .
* each leaf node is
  + the path from to , where
  + is the inner branches in this path
  + , with the that each branch as a probability (**a conditional probability given we know where the branch started**).
* for is the product of branch probabilities from root to final leaf

**this is for the leaf level**

**this is for the internal nodes** the joint density of has the form of:

* is the number of children of

that each component in the product corresponds to a interior node in the tree, which is a Dirichlet multinomial distribution based on the accumulated counts along branches of given node.

given

Tips: 1. when , each has an independet variance 2. two components $p\_{v\_i}, p\_{v\_j} $ with in the same subtree indexed by are correlated, due to the same ancestor of .

### Dirichlet Tree Multinomial Regression Model

a regression of on covariates

for each and we express as a linear combination of the covariates.

### Regularized Likelihood Estimation

parametrization for Dirichlet Tree Multinomial Regression Model

number of time the number of tree branches

### Algorithm of accelerated proximal gradient method

approximate at