Thesis Outline

Casey Sakamoto

5/15/2020

# Latent Class Model in the Airway Microbiome

## Introduction

### Cystic Fibrosis Relevance

Cystic Fibrosis (CF) is the most common life-threatening autosomal recessive disease in US, affecting 1 in 4000 newborns in US, with a higher rate in some European countries. (Farrell 2017) CF Pulmonary disease begins early and can progress rapidly without early identification. (Goetz 2019) CF leads to airways that are vulnerable to chronic bacterial infections, requiring repeated hospital visits, IV antibiotics, and additional care. (CF org) Subjects with CF have varied disease progression and a better understanding of airway infections early in life may help us better understand this variability.

It is therefore important to understand the microbiome of airways within CF patients so that we can understand discrepancies in subject’s microbiome composition as they age. Furthermore, identifying phenotypes based on airway microbiome community composition may help us learn about the similarities in the microbiome change over time; are the observable differences in subject microbiomes due to inherent subject variability, or do subjects start with similar microbiome compositions that diverge over time? Currently, there are multiple ways to cluster and build these phenotypes (see highlighted articles section), however the complexity of the data collected from the microbiome necessitates the use of statistical methods which are often more complex than that of other studies.

### Microbiome background

The microorganisms living both inside and on the human body outnumber human somatic and germ cells tenfold. The microbiome provides traits to the human system that humans did not need to evolve on their own such as nutrient metabolism within the gut microbiomes. We can use the characteristics of the human microbiome to identify key differences between members of a family, community, or across different environments and populations. Collecting microbiome data (with the goal of a “random sample”) requires the consideration of aspects such as the location and quantity of samples. The difficulty of collecting such samples varies with the system and populations of interest, and in many cases the number of samples is often limited. (Turnbaugh 2007)

Sequencing of a specimen (such as a sputum sample) provides an estimate of the abundance of different operational taxonomic units (OTU’s), which are microbial sequences clustered by sequence similarity. These sequences are then referenced (often with a database) to get an abundance of different species within the sample. It is important to note that this is an estimate and not a true abundance. This data is therefore a count of different species within a sample, and relative abundances are often used to compare populations. (Mandal 2015) This example highlights a potential pitfall of using counts rather than relative abundance to compare samples:

If we sample 100 dogs from dog parks A and B, and get 10 and 20 corgis, it is reasonable to estimate that 10 and 20% of dogs in each park are corgis, however we can’t conclude that there are more corgis in park B; if park A has 1000 dogs and park B has 200, then park A would have 100 corgis and park B would have 40.

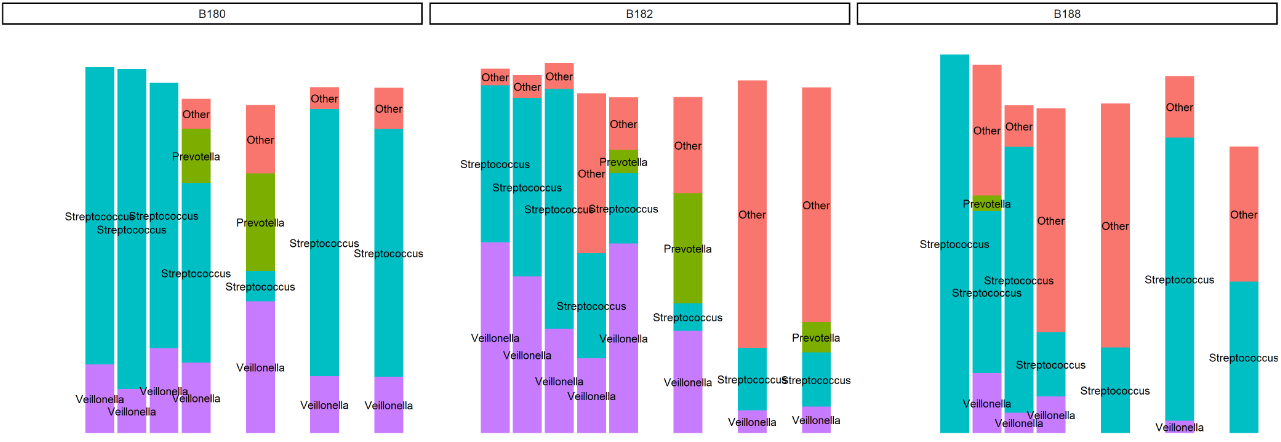
It is important to note that relative abundances sum to 1, which could lead to incorrect results when applying standard statistical methods such as Pearson’s Correlation, t-tests, ANOVA, and linear regression. The lack of measurement precision leads to additional difficulty in drawing conclusions from analysis results.

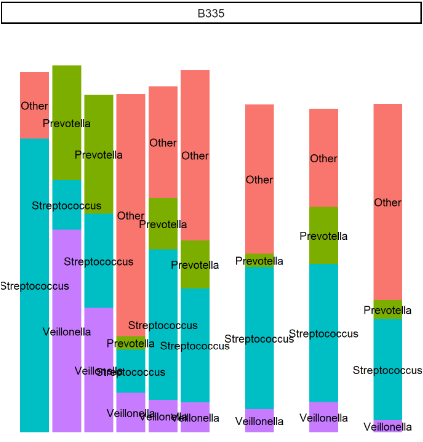
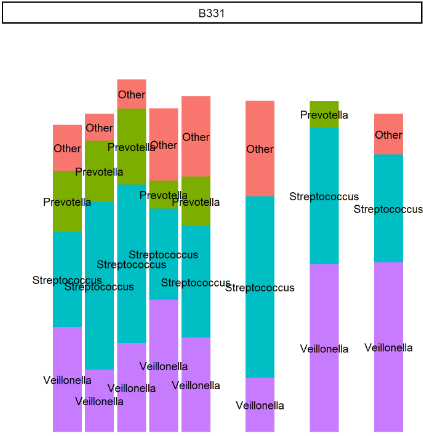
### Motivating example [BONUS study]

Goetz et al. (2019) analyzed data from 231 infants within 28 centers in the US in the Baby Observational and Nutritional Study (BONUS) which was comprised of clinical, medication, symptoms, culture, and chest radiographs data collected repeatedly throughout the first year of life. 1053 samples in total were collected, and each subject had around 4-6 samples. Age was measured in months, both continuous and “rounded;” if there was a patient with months 4.7 and 5.4 recorded, these months were rounded to 5 and 6. 194 infants (84%) developed a CF specific pathogen, and there was a relatively low number of taxa within the collected samples. (we could categorize into groups).

### Why I am looking at latent class analysis

The application of latent class analysis to the microbiome is still relatively novel, especially relating to the airway microbiomes, as there has not been a lot of recent applications, especially incorporating mixed components. Several papers have emphasized the utility of building latent variable models in the context of examining the true bacterial composition of a sample (Wu, Berkow 2013; daSilvaSolca 2014; J Fu 2015; Sundarenson 2018). Specifically, (Wu,Berkow 2013) mentions that RMANOVA is a special, more constrained case of latent variable structural equation modeling and that latent variable models provide a more flexible, better fit. This type of modeling has few assumptions on the data which makes it appealing for use on the typically messy microbiome data. The main assumption is that data are normally distributed, however there are several transformations commonly used on compositional data to meet this requirement. (Aitchison 1986) Several studies have also used a similar Bayesian method to get around the normality assumption.

The following stacked barcharts are the microbiome compositions of 5 subjects from the BONUS Study; The different trajectories of the groups within subjects (“consistent” Veillonella in B180, B331 vs “decreasing” in B182, B335 etc) could indicate the presence of these latent classes.



## Thesis Questions

I am interested in identifying airway microbiome phenotypes in infants with CF.

Specifically, my main questions are:

1. how many latent classes are there/can we build these classes using LC analysis?

* In Brandie’s work there were 4 clusters from random forest – point of reference

1. When we build these classes what covariates are associated with our clusters?

* Start with just age

1. [Optional] How does the Latent Class model compare with 2-stage models

* If clinical findings for (1) and (2) not interesting

## Methods

I am looking to build a multivariate latent class mixed model, which can adjust for fixed effects and random effects due to within-subject correlation to identify phenotypes (# latent classes).

### Building a Latent Class Linear Mixed Model (Proust 2017 for all math sections)

#### What is a Latent Variable? (bollen 2002)

A latent variable is a variable for which there is no sample realization for at least some (sometimes all) observations in a sample. With regular observed random variables, a sample will contain realizations. For example, we can’t measure self-esteem (latent), but we can use the compiled answers on a questionnaire (observed r.v.) to glean information about it. Wu, Berkow (2013) defines the true bacterial composition in a sample as a latent variable. However, several measurement platforms are used to measure this composition. Latent variable analysis was used to identify the most reliable measurement platform.

I am interested in treating the phenotypes of subjects as a latent variable, and using the measured composition of their microbiome communities (specifically their trajectories) to build these phenotypes.

### how does this compare to other methods

The latent class mixed model is a sub-genre of structural equation modeling (SEM), and very similar to cluster analysis. We are interested in identifying unmeasured clusters (phenotypes) via the trajectories of the microbial composition of the airways of the subjects. With this method we are able to incorporate a linear mixed model framework into the modeling of our phenotypes to adequately fit longitudinal microbiome data, where regular mixed models may be inappropriate.

#### LMM recap

Recall that for a sample of subjects with repeated measures , the outcome value for subject at the measurement at time , the linear mixed model is as follows:

,

Where is our vector of fixed effects and is our vector of random effects. Trajectory shapes in and can be “of any type such as polynomial, specifically designed to fit, or approximated using basis of I-Splines.” The random effects vector has a zero-mean MVN distribution with unspecified covariance structure. Measurement errors are independent Normal errors with variance . The process is a zero-mean Gaussian stochastic process or a stationary process with parameter (note: this is included in semiparametric lmm, not regular parametric lmm from BIOS6643, and is used to model serial correlation). In the LMM, we assume Normally distributed deviations (measurement errors, random effects, correlated errors) and constant covariate effects across time. However, these assumptions do not hold for many outcomes, especially in the microbiome where an outcome is a count.

#### Latent Process Mixed Model

Thus, Proust defines a family of models, the “latent process mixed models” as “separating the structural model that describes the quantity of interest (latent process) according to time and covariates from the measurement model that links the quantity of interest to the observations.”

Define the latent process as a standard linear mixed model without measurement error:

In order to account for different types of longitudinal markers (nonnormal), a link function between the latent process and the outcome at time :

Where are the parameters to the link function which transforms outcomes to Normally distributed to fit framework of standard LMM. It is important to mention that the package has several options to transform the outcome: a linear transformation to a gaussian framework, rescaling Y to (0,1) and using the beta CDF, and using a basis of quadratic I-splines. The latent process mixed model constrains the location of and the scale .

#### Multivariate Longitudinal Markers

The previous model is extended to the multivariate case involving multiple longitudinal markers (eg. different bacteria taxa) with function *multlcmm()* , and the latent process is the common factor underlying the markers. In this case, we can include covariates with marker-specific effects called contrasts (where the sum of these coefficients = 0). Each marker relationship with underlying quantity of interest is modeled through its own link function similar to previous section. Fitting this model has similar constraints to those in the previous section; although this also requires a random intercept, and no mean intercept is allowed in the structural model.

### Latent Class Mixed Models

Assume a heterogenous population with latent classes of subjects, where each subject belongs to one class. Then is a discrete random variable that equals when subject i belongs to latent class . P() is a class-specific probability, but can also be described using a multinomial logistic regression model with time-independent covariates .

For a Normal outcome in class , the linear mixed model is:

,

Where X are fixed effects (can be split into common and class-specific), Z are individual random effects, w is a normal autocorrelation process, and is random noise.

The extension of this framework to a latent process model becomes:

The distribution of u is class specific where B is an unspecified VCOV matrix, and w is a class-specific coefficient to allow class-specific individual variability. This model assumes the population heterogeneity only affects the underlying latent process of interest. The parameters to be estimated are : Fixed effects, Random effects, variance of w, variance of the errors, parameters for beta transformation (link for outcome to latent process, p(c = g), w)( quite a few parameters).

### Estimation of parameters and likelihoods

The individual contribution to the likelihood of a lcmm is:

Where is a random variable that equals when the subject belongs to latent class , and

A subject’s individual contribution is obtained by replacing with the appropriate parameters. In the lcmm package, an extended Marquadt algorithm (Newton-Raphson family) will be used to maximize the log-likelihood. The algorithm works by updating the vector of parameters until convergence using the following:

note: denotes iteration

Where is a diagonal-inflated Hessian (helps ensure positive definiteness) and is the gradient of the log-likelihood at the th iteration. Convergence of the log-likelihood is based on parameter stability, log-likelihood stability, and the size of the derivatives. All three must satisfy convergence criteria (threshold which is default ). Having multiple criteria is important because the shape of the log-likelihood in lcmm can be relatively flat in areas of the parameter space.

## Proposed Analysis

**Data Manipulation**

I will be looking to split the sample data into groups: Streptococcus, Prevotella, Veillonella, and Other are the preliminary groups (suggested by Brandie). The Streptococcus genus made up 50.9% of the total sequence counts, Veillonella made up 12.3%, Neisseria made up 7.6%, and Prevotella made up 6.7%. It could make sense to make 5 groups (Streptococcus, Veillonella, Neisseria, Prevotella, and Other), or substitute Neisseria for Prevotella as one of the groups. These were the only groupings with counts over 10 million, with the next most counted genus being Gemella, and class Bacilli with ~ 5 million counts each.

In order to fit a latent class mixed model, I will normalize the outcomes; for compositional data, two common transformations are the centered logratio transformation (CLR) and the isometric logratio transformation (ILR).

The CLR transform of vector x is:

The ILR transform of a vector x is:

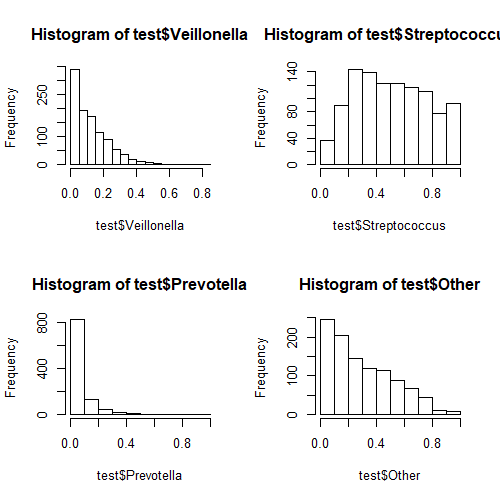
,

where *VT* is an orthonormal basis of the clr-plane.

While the CLR transformation is more simple to compute and interpret, the resulting parts sums to 0, which may lead to singularity issues in model fitting (However, this may not be an issue depending on whether we use all parts of the composition). Because of this the ILR (which does not run into singularity issues at the cost of interpretability) will also be considered if the CLR transformation is inadequate. A potential downside of using the ILR is that the transformation maps the D parts into D-1 parts, where there is no 1-1 relation between the original parts and the transformed variables.

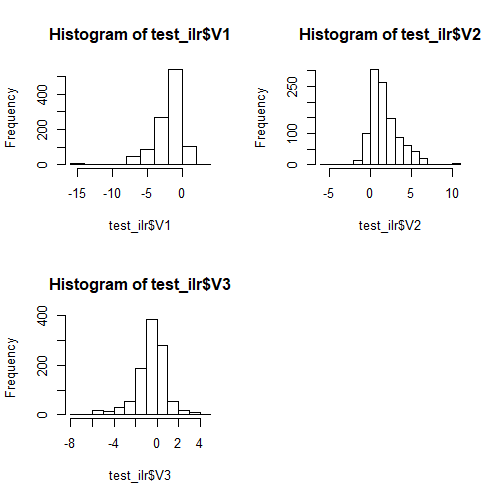
**CLR**

The CLR transformed variables do not look particularly normal, which could lead us to search for other transformation methods:



**ILR**

The ILR transformed variables look a lot better, but still slightly skewed.



**Model**

Once outcomes are normalized, primary models with only age will be fit. A simple form, such as linear or a small order polynomial for age will be used. Similarly, a random intercept for subjects will be included in the model. We could also consider models with specific covariance structure form (such as AR(1) or a spatial structure) as well. Once the functional form of the fixed and random components are decided, models with different numbers of latent classes (1 – 10) will be fit.

It is important to mention that when fitting models with multiple latent classes, initial values for the parameters Θ need to be specified. These specifications can play in a role in the convergence of the model. In the package documentation, these values are specified from the model with only one latent class. This is done internally and may increase estimation time, so an alternative could be a random draw from the model with one latent class, where initial values are generated from the asymptotic distribution of the MLE of this model.

The model with the optimal number of latent classes (# of phenotypes characterized) will be selected with the Bayes Information Criterion (BIC) because of the relatively high number of parameters estimated.

Furthermore, once a model is selected, a second model will be fit to assess covariate associations.

**Output/ Results**

A “table 1” could be useful to include where it describes the data.

A table describing the models fit and their model fit criteria (BIC, number of parameters estimated, posterior latent class membership) could illustrate how we chose the model, and how it compares to other models considered.

Included in the lcmm package is the function postprob() which classifies the subjects into the latent classes for which they have the highest posterior class-membership probability. This would show the distribution of the subjects among latent classes, as well as providing the mean of the posterior probabilities of belonging to each class. This could be used to assess a goodness of fit relating to the number of latent classes specified in the model by giving insight into the ambiguity of subject classification.

One way to characterize the phenotypes is by plotting the trajectories of the latent classes for each outcome. This could describe how the changes over time of each outcome differs between phenotypes. Additionally, it could be similarly beneficial to examine stacked barcharts or a similar figure for specific times (such as baseline, or the final measured time).

Finally, a table describing the associations between covariate and outcomes could be a concise way of answering the second main question.

Residuals can be plotted to assess normality of outcome.

# sources

### Highlighted articles using latent class models

Bacharier (2019) - phenotypes & mixed models with trajectories for airways in children

Eun Lee (2017) - identified phenotypes in children

L Xu (2017) - bayesian LV model to jointly model mult phenotypes in longitudinal family studies

Proust-Lima (2017)

JS Son (2015) - looked at lv SEM in fecal microbiome in children with ASD, built classes

### other sources

Pulmonary findings in infants with cystic fibrosis during the first year of life: results from the Baby Observational and Nutrition Study (BONUS) cohort study. Goetz D 2019 - data set for our paper

Diagnosis of Cystic Fibrosis: Consensus Guidelines from the Cystic Fibrosis Foundation (2017) - background on cf

Characteristics and outcomes of oral antibiotic treated pulmonary exacerbations in children with cystic fibrosis Farrell (2018) - background

<https://www.cysticfibrosis.org.uk/what-is-cystic-fibrosis/how-does-cystic-fibrosis-affect-the-body/symptoms-of-cystic-fibrosis/lungs> - background on cf

The Human Microbiome Project Turnbaugh 2007 - background on microbiome analysis of composition of microbiomes mandal 2015- background on microbiome data

<http://www.john-uebersax.com/stat/faq.htm#otherm> - cites some papers (to be read/ cited later) but gives a nice overview of lca

### some nice figures sources

spycher, silverman 2008 shows some barcharts for their phenotypes, could be interesting if we want to look at a specific time

sakai, boardman 2010 shows a trajectory of their phenotypes which could be a very useful plot for each of our outcomes

Jackson, sher 2008 show a nice latent growth curve approach, which is different from lca, but has some nice comparisons of trajectories (and include a mixture model for comparison in a previous 2005 paper)