Thesis Outline

Casey Sakamoto

3/3/2020

# Latent Class Model in the Airway Microbiome

## Introduction

### Cf relevance

Cf is most common life-threatening autosomal recessive disease in US, affecting 1 in 4000 newborns in US, and a higher rate in some european countries (Farrell 2017)

CF pulmonary disease begins early, and can progreess rapidly without early identification. (goetz 2019)

CF leads to airways vulnerable to chronic bacterial infections, requiring repeated hospital visits, IV antibiotics, and additional care (CF org)

It is therefore important to understand the microbiome of airways within CF patients so that we can create treatment plans for patients with different airway bacterial communities.

Furthermore, identifying phenotypes based on airway microbiome community composition may help us specialize treatment for patients with these distinct phenotypes.

Currently, multiple ways to cluster and build these phenotypes (see highlighted articles section), and it is difficult to methods/truth

### Microbiome background

#### Turnbaugh 2007

The microorganisms living inside/on humans outnumber human somatic and germ cells (body/reproductive) tenfold

The microbiome provides traits to the human system that humans did not need to evolve on their own (i really like figure 1 in this paper)

We can use the microbiome to identify key differences between members of a family, community, or across different environments

Microbiome data is important; to answer any scientific question, we need sample data, and scales(temporal: interval between measurements, spatial: where are we measuring, and how difficult is it to “randomly sample” the system[depending on what organ it could be from micrometers to meters], what populations are we interested in)

#### analysis of composition of microbiomes mandal 2015

Sequencing of a specimen (such as a sputum sample) provides an estimate of the abundance of different OTU’s (operational taxonomic units; clustered microbial sequences clustered by sequencce similarity)

Important to mention this is an estimate and not a true abundance; the lack of precision leads to difficulty in analysis

In addition, relative abundances are often used to compare populations;

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.

relative abundances sum to 1, which could lead to incorrect results when applying standard statistical methods (pearson’s corr, t-test, anova, linear regression)

### Motivating example [BONUS study]

(Goetz 2019) analyzed data from the Baby Observational and Nutritional Study (BONUS): 28 centers in US which has clinical, medication, symptomes, culture, and chest radiographs data collected repeatedly throughout the first year of life (; around 4-6 samples/subj; 1053 samples total)

time measured in age (months) both continuous and “rounded” (if there was a patient with month 4.7 and 5.4, these months were rounded to 5 and 6)

194 infants (84%) developed a CF specific pathogen

relatively low # of taxa (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; not a lot of recent application, especially incorporating mixed components, to airways.

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).

(Wu,Berkow 2013) mentions that RMANOVA special, more constrained case of LV SEM and that LV models more flexible, better fit

“All we need” is a normally transformed outcome to use this method

Several studies used a similar method using Bayesian LV models to get around normality assumption

## Thesis Questions

I am interested in the similarities in phenotypes during a pulmonary exacerbation (PEx) in children.

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 given 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 a latent variable as the true bacterial composition in a sample. However, several measurement platforms are used to measure this composition. Latent variable analysis used to identify the most reliable measurement platform.

### how does this compare to other methods

Very similar to cluster analysis; the goal is to discover groups based on data; this is the goal of my analysis

sub-genre of cluster?

what we are interested is measured imperfectly (latent)

similar to factor analysis, but factor analysis covers structure of variables (correlations), while lca covers structures of the structure of the cases themselves

what’s nice about this method is its ability to incorporate components of a multivariate (generalized) linear mixed model into the latent classification #### 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.

in package: linear transformation to gaussian framework, rescale Y to (0,1) and use beta CDF, basis of quadratic I-splines

The latent process mixed model constrains the location of and the scale

#### latent process mixed model for multivariate longitudinal markers

The previous model extends to the multivariate case involving multiple longitudinal markers (eg. different bacteria taxa), and the latent process is the common factor underlying the markers.

In this case, we can include covariates with marker-specific effects called contrasts (sum of coefficients = 0)

Each marker relationship with underlying quantity of interest is modeled through its own link function similar to previous section.

similar constraints to 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 .

the probability of is a class-specific probability, but can also be described using a multinomial logistic regression model with me-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.

extended to a latent process model:

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: 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 beccause the shape of the log-likelihood in lcmm can be relatively flat in areas of the parameter space.

## Proposed Implementation

I will need to decide before fitting the model which covariates I will include, and what kind of random effects to include.

For primary models, we will fit only using age to see if we can build classes with the model.

In order to fit this model, we need our outcomes to be normalized. For compositional data, two common transformations are the centered logratio transformation and the isometric logratio transformation. 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. Because of this the ILR (which does not run into singularity issues at the cost of interpretabilty) will be used.

# 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

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