Insights into Cystic Fibrosis

Jamie Morton, University of California, San Diego

Review on Regression

Coefficients

$$y = \beta_0 + x_1\beta_1 + \cdots + x_p\beta_p + e$$

Response
(Dependent variable)

(i.e. height)

Covariates
(Independent variable)

(i.e. foot size, hand size, …)

Error
(Residual)

Review on Regression

$$Y = X \quad B + e$$

$$\underset{\substack{\text{n x 1} \\ \text{Response}}}{\begin{bmatrix} \boldsymbol{y_0} \\ \boldsymbol{y_1} \\ \dots \\ \boldsymbol{y_n} \end{bmatrix}} = \underset{\substack{\text{n x p} \\ \text{Covariates}}}{\begin{bmatrix} 1 & \boldsymbol{x_{11}} & \dots & \boldsymbol{x_{1p}} \\ \dots & \dots & \dots & \\ 1 & \boldsymbol{x_{n1}} & \dots & \boldsymbol{x_{np}} \end{bmatrix}} \underset{\substack{\text{p x 1} \\ \text{Coefficients}}}{\begin{bmatrix} \boldsymbol{\beta_0} \\ \boldsymbol{\beta_1} \\ \dots \\ \boldsymbol{\beta_p} \end{bmatrix}} + \underset{\substack{\text{n x 1} \\ \text{Residuals}}}{\begin{bmatrix} \boldsymbol{e_0} \\ \boldsymbol{e_1} \\ \dots \\ \boldsymbol{e_n} \end{bmatrix}}$$

n = number of measurements

p = number of variables measured

Assumptions

- Over-determined system (n>>p)

- Independence between measurements

- Compositionality

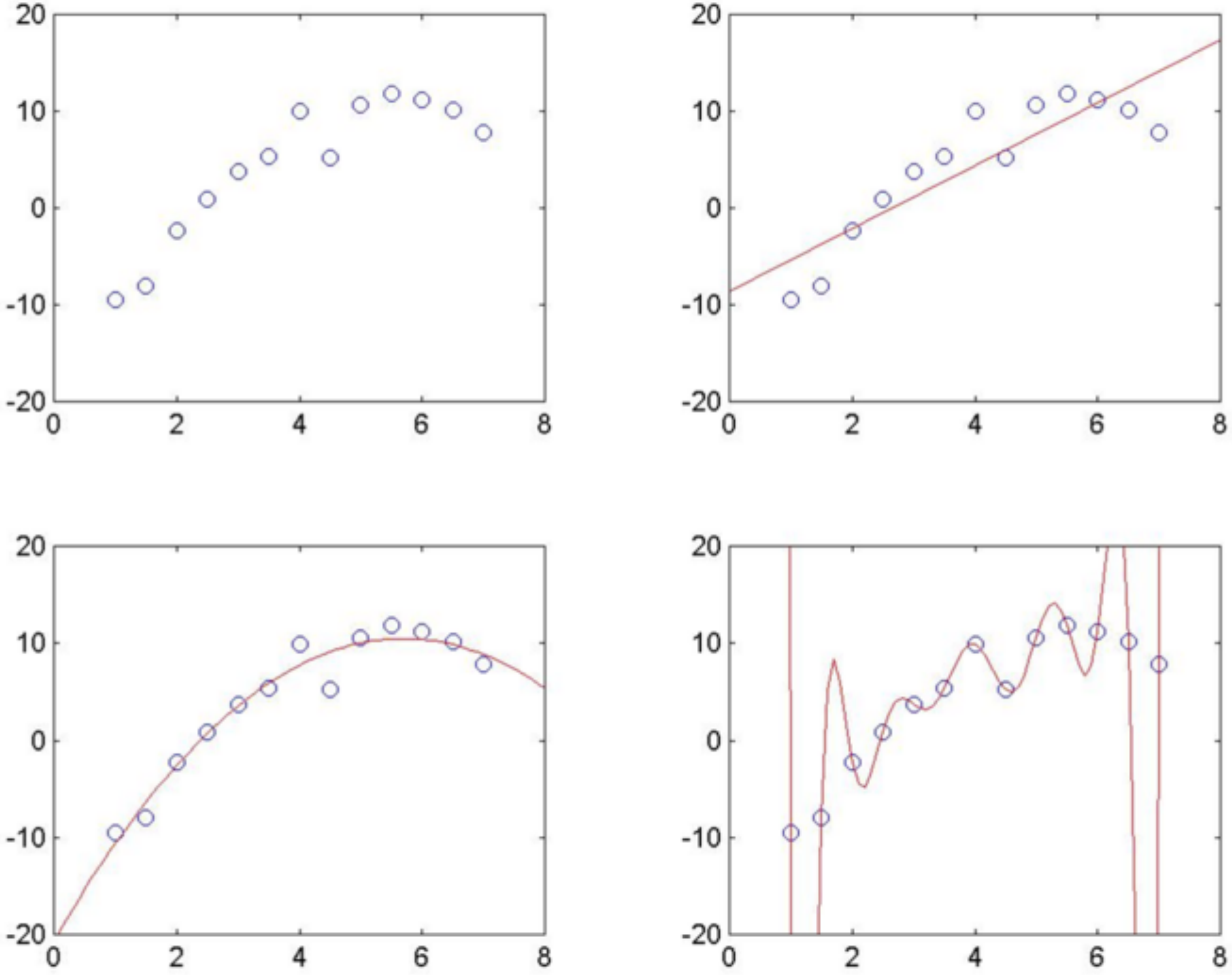
Assumptions

- Over-determined system (n>>p)

- Independence between measurements

- Compositionality

Occams Razor



n = 13
p = 1

n = 13
p = 2

n = 13
p = 13

High dimensional variables

- Genetic data
  - Thousands of variables
  - Hundreds of samples

- Two solutions

  1. Regularization – wisely choose a subset of variables

  2. Multivariate response

High dimensional variables

- Genetic data
  - Thousands of variables
  - Hundreds of samples

- Two solutions

1. Regularization – wisely choose a subset of variables

2. Multivariate response

Univariate response

$$
Y = X \; B
$$

$$
\begin{bmatrix} y_0 \\ y_1 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ \dots & \dots & \dots & \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_p \end{bmatrix}
$$

n x 1	n x p	p x 1
Response	Covariates	Coefficients
(BMI, age, sex, …)	(gene abundances)	

n = number of measurements

p = number of variables measured

Only one variable at a time!

Multivariate response

Y = X B

$$
\begin{bmatrix} y_{00} & y_{01} & \dots & y_{0D} \\ y_{10} & y_{11} & \dots & y_{1D} \\ \dots & \dots & \dots & \dots \\ y_{n0} & y_{n1} & \dots & y_{nD} \end{bmatrix}
=
\begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ \dots & \dots & \dots & \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}
\begin{bmatrix} \beta_{00} & \beta_{01} & \dots & \beta_{0D} \\ \beta_{10} & \beta_{11} & \dots & \beta_{0D} \\ \dots & \dots & \dots & \dots \\ \beta_{p0} & \beta_{p1} & \dots & \beta_{pD} \end{bmatrix}
$$

n x 1

Response

(gene abundances)

n x p

Covariates

(BMI, age, sex, …)

p x D

Coefficients

Can encode categorical variables

n = number of measurements

p = number of covariates measured

D = number of variables measured

Advantages

- Effectively avoids over parameterization
    - n >> p
    - But always try to run cross validation to confirm
        - Train model on subset of measurements
        - Try to predict the remaining measurements
- Build models with many covariates

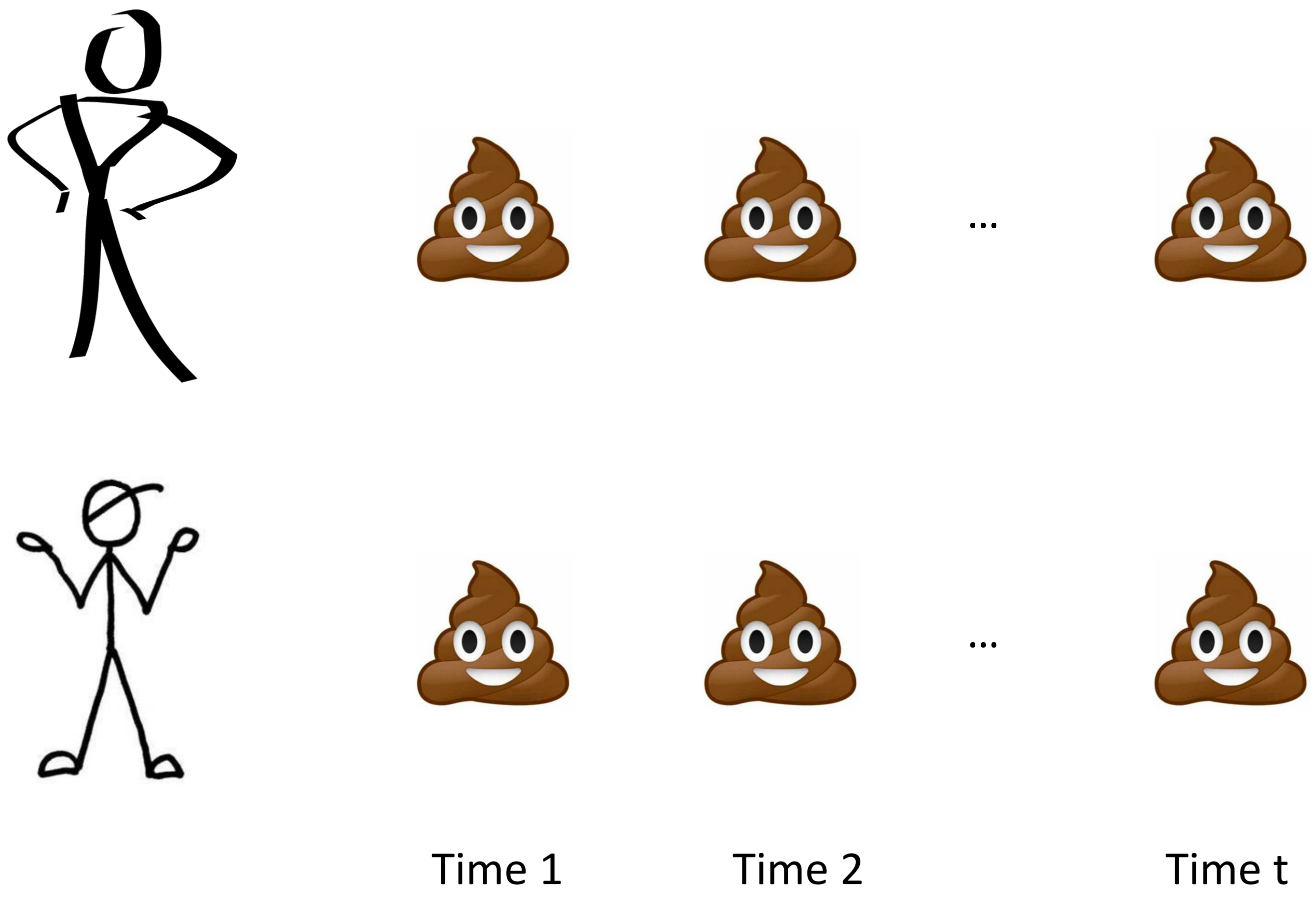
Assumptions

- Over-determined system (n>>p)

- Independence between measurements

- Compositionality

Independence



- Independence is violated
- Samples depend on person
  - Drawn from different distributions

Linear Mixed Effects Models

$Y = X\beta + e$ Linear Regression

$Y = X\beta + Z\mu + e$ Linear Mixed Effects Model

Fixed effects

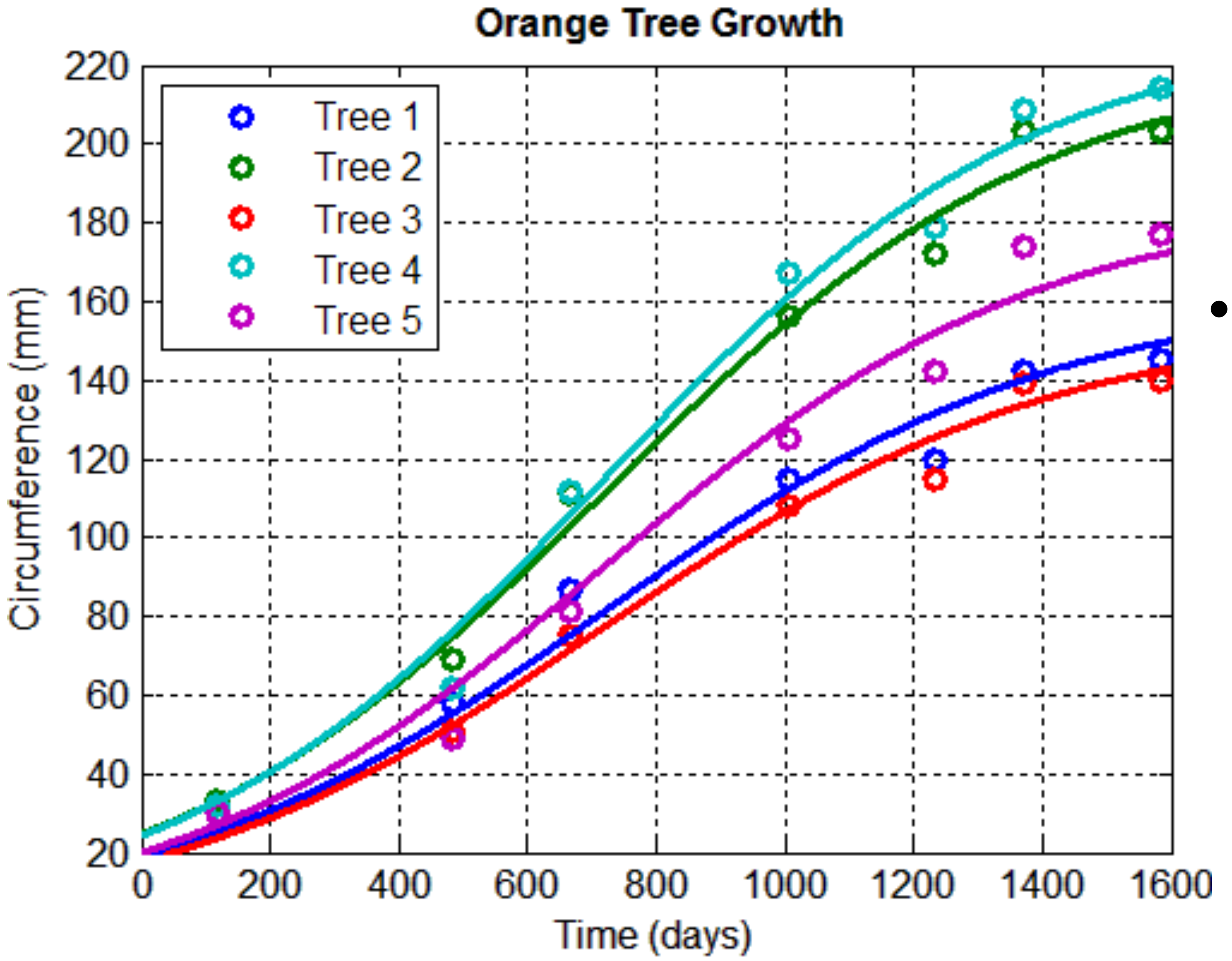
(i.e. sex)

Random effects

(i.e. test score)

Fixed effect: constant value
Random effect: value varies

Linear Mixed Effects Models



- Fit multiple regressions
  - One per individual

Assumptions

- Over-determined system (n>>p)

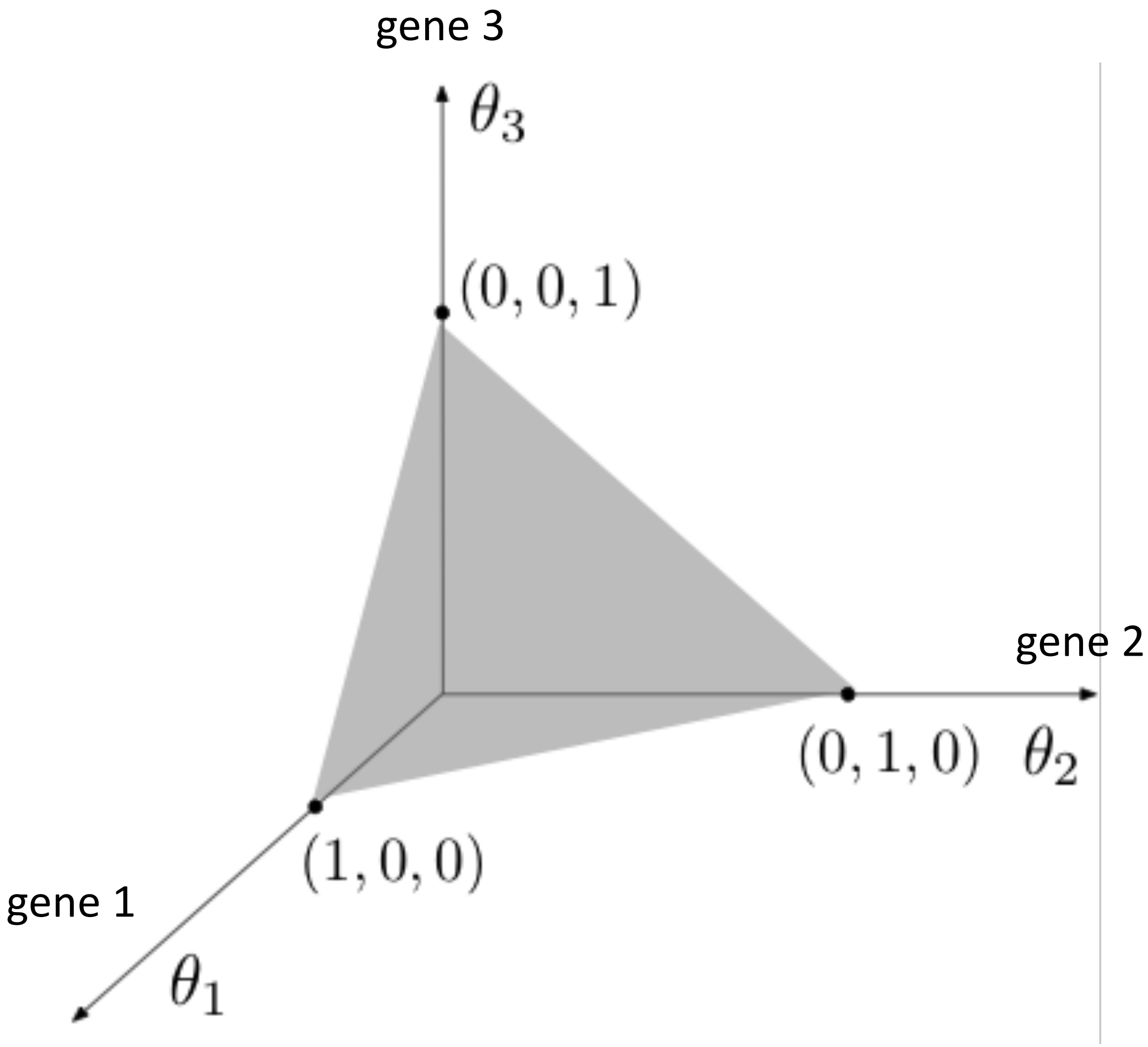
- Independence between measurements

- Compositionality

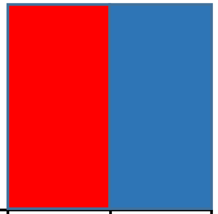
Compositionality

- We're dealing with proportions

- Thus all of the our samples live in the simplex

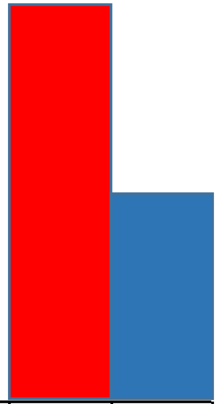


Compositionality

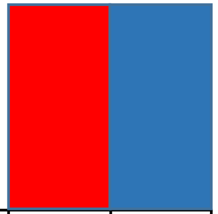


Proportions	1/2	1/2
Counts	100	100

Red doubled

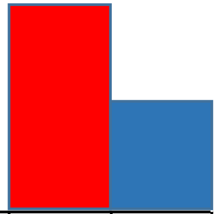


Proportions	2/3	1/3
Counts	200	100



Proportions	1/2	1/2
Counts	100	100

Blue halved



Proportions	2/3	1/3
Counts	100	50

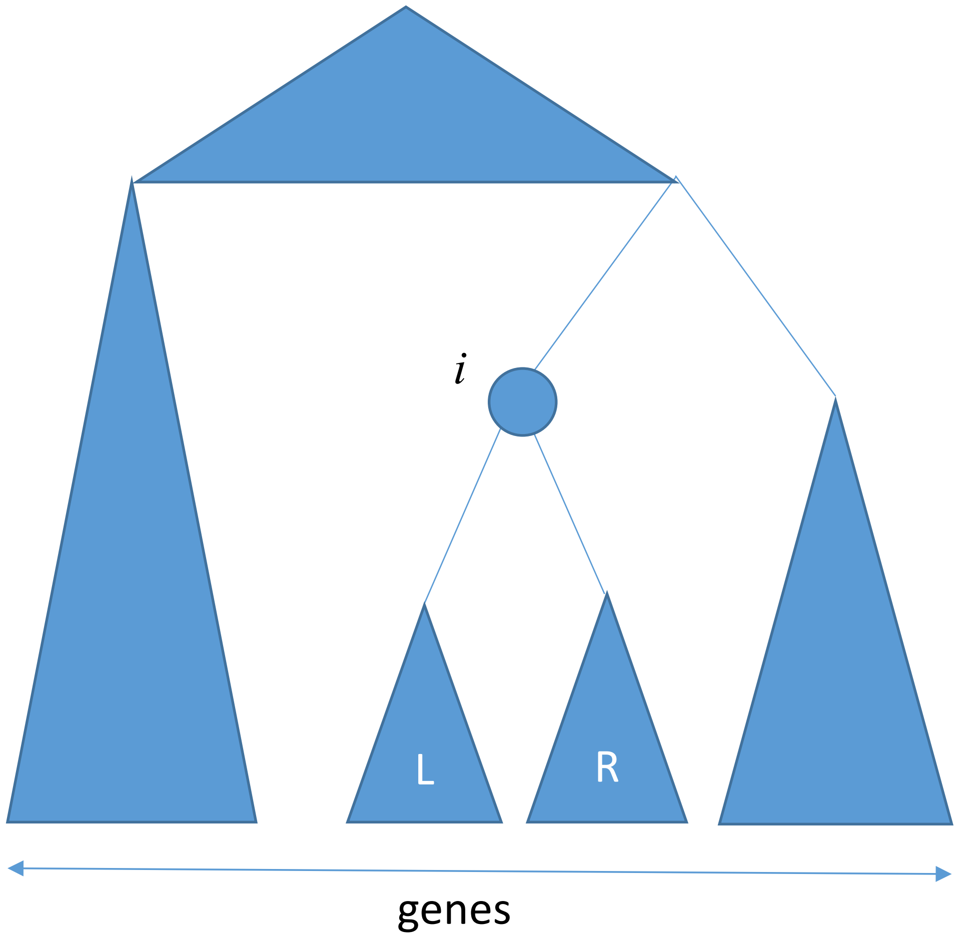
Time point 1

Time point 2

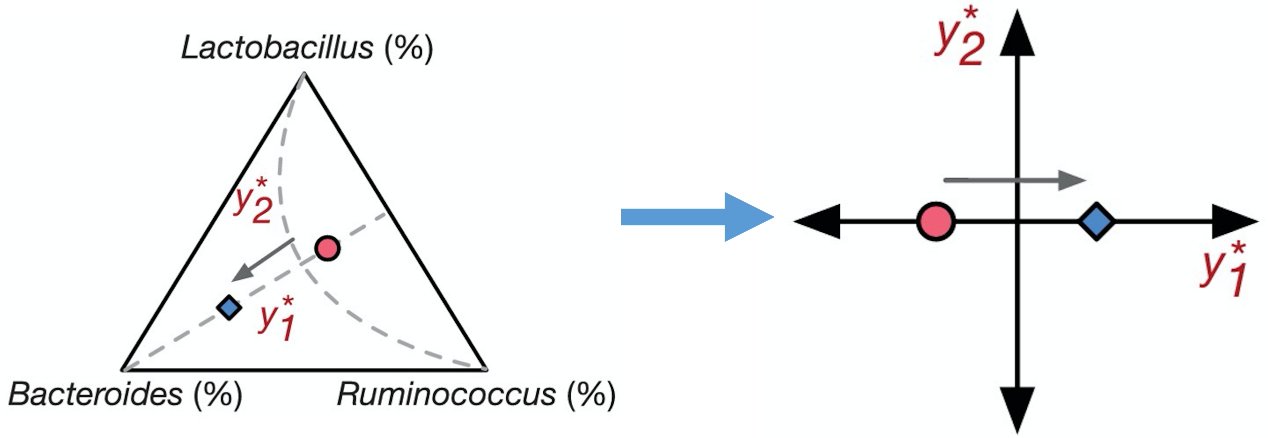
This breaks regression

- Cannot predict who's growing/dying
- Solutions must be restricted to the simplex

The isometric log ratio transform



$$ilr: S^D \rightarrow \mathbb{R}^{D-1}$$



$$b_i = \sqrt{\frac{|i_L||i_R|}{|i_L| + |i_R|}} \log\left(\frac{g(i_L)}{g(i_R)}\right)$$

$$g(x) = \sqrt[k]{x_1 \dots x_k}$$ Geometric mean

Silverman et. al

Recap

- Multivariate response resolves Occam's razor
- Linear mixed effects models resolves sample dependence issue
- ILR transform resolves compositionality problem

Lesson Plan

- Perform ILR transform
  - Build tree based on pH hierarchical clustering
  - i.e. group organisms that live in similar pH together

- Build Multivariate Response Linear Mixed Effects Model

$$Y = X\beta + Z\mu + e$$

Microbial proportions

pH

Patient id

Software

https://biocore.github.io/gneiss/

https://qiime2.org/

http://scikit-bio.org/