Compositional Models for Microbiome

HMP2Data package

https://hmpdacc.org/ihmp/

https://github.com/dozmorovlab/HMP2Data

Workshop:

BiocWorkshops2019

Compositional data

Vector of proportions

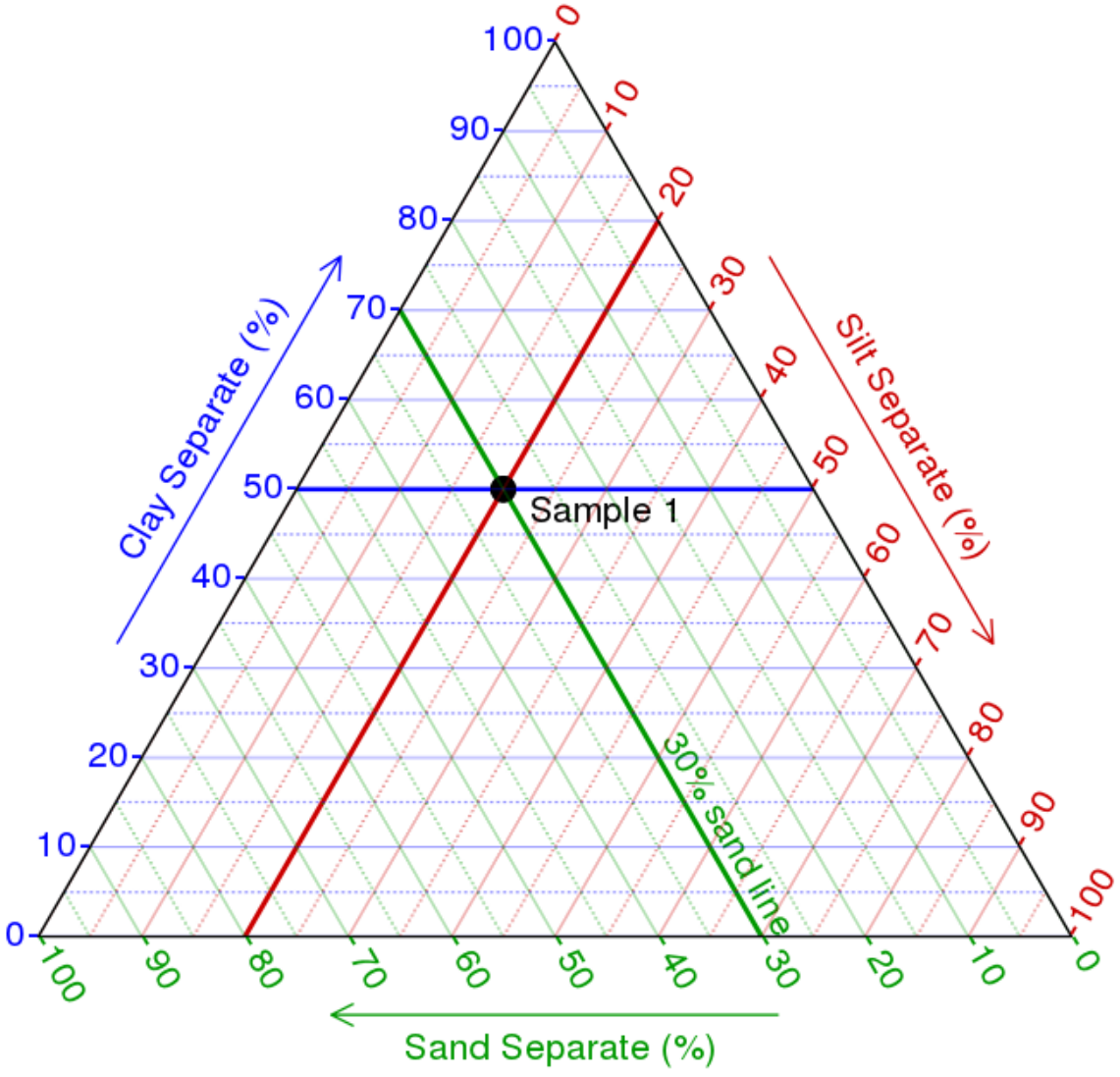
$$z = (z_1, \ldots, z_k)^T, z_i > 0, \sum_{i=1}^{k} z_i = 1, z \in \Delta^{k-1}$$

Closure operation: $\mathcal{C}[z_1, \cdots, z_k] = [\frac{z_1}{\sum_{j=1}^{k} z_j}, \cdots, \frac{z_k}{\sum_{j=1}^{k} z_j}]$

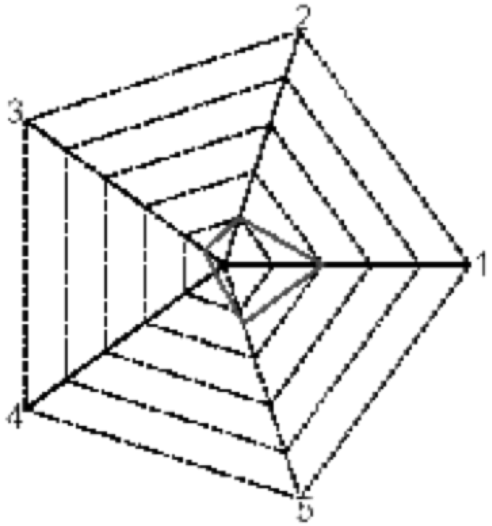
- Composition of rock samples.
- Composition of nutrient groups in diet
- Composition of air pollutions

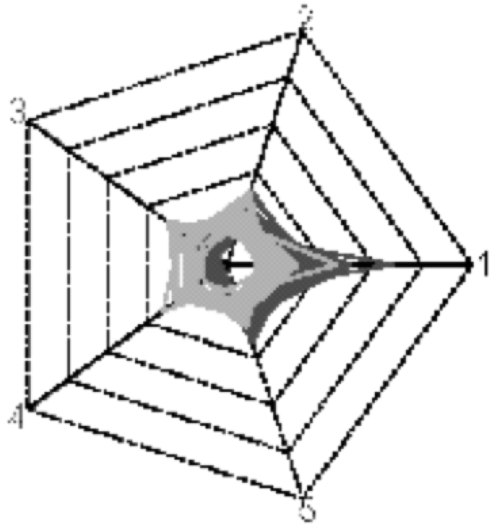
A subcomposition z_s with s parts is obtained via the closure of a subvector $[z_{i_1}, z_{i_2}, \ldots, z_{i_s}]$ of z.

Triangle plot



Spider plot





$(0.4, 0.2, 0.1, 0.05, 0.25)$

Proper CoDa analysis

- scale invariance: multiple the vector with a scalar doesn't change anything
- permutation invariance: the order of the parts should be irrelevant
- subcompositional coherence: studies performed on subcompositions should not stand in contradiction with those performed on the full composition

Algebra for compositions

- Perturbations: For $\xi, \alpha \in \Delta^{k-1}$

$$\xi \oplus \alpha = \left(\frac{\xi_i \alpha_1}{\sum_{i=1}^k \xi_i \alpha_i}, \ldots, \frac{\xi_k \alpha_k}{\sum_{i=1}^k \xi_i \alpha_i} \right)$$

  The composition $\mathsf{e} = (\frac{1}{k}, \ldots, \frac{1}{k})$ acts as zeros so that $\xi \oplus \mathsf{e} = \xi$

- Let $\xi^{-1} = (\frac{1}{\xi_i} \ldots, \frac{1}{\xi_k})$, then $\xi \oplus \xi^{-1} = \mathsf{e}$
- $\xi \ominus \eta = \xi \oplus \eta^{-1}$

Power

Let a be a scalar.

$$\xi \otimes a = \left(\frac{\xi_1^a}{\sum_{i=1}^k \xi_i^a}, \ldots, \frac{\xi_k^a}{\sum_{i=1}^k \xi_i^a} \right)$$

Ex: Aitchison (1986) regression of sand, silt and clay in rock composition.

$$x = \xi \oplus [\log d \otimes \beta] \oplus p$$

Vector space structure

- commutative group structure of $S^k, \oplus$
  - commutativity: $\xi \oplus \eta = \eta \oplus \xi$
  - associativity: $(\xi \oplus \eta) \oplus \epsilon = \eta \oplus (\xi \oplus \epsilon)$
  - Inverse: $\xi \ominus \xi^{-1} = \mathsf{e}$ and $\xi \oplus \eta^{-1} = \xi \ominus \eta$
- properties of powering
  - associativity $a \otimes (b \otimes \xi) = ab \otimes \xi$
  - distributivity 1 $a \otimes (\xi \oplus \eta) = (a \otimes \xi) \oplus (a \otimes \eta)$
  - distributivity 2 $(a + b) \otimes \xi = (a \otimes \xi) \oplus (b \otimes \xi)$

Inner product space

- $(\Delta^{k-1}, \oplus, \otimes)$ is a complete inner product space, with

$$\begin{aligned}\langle \xi, \eta \rangle &= \sum_{i=1}^{k} \log \frac{\xi_i}{g(\xi)} \log \frac{\eta_i}{g(\eta)} \\ &= \frac{1}{2k} \sum_{i=1}^{k} \sum_{j=1}^{k} \log \frac{\xi_i}{\xi_j} \log \frac{\eta_i}{\eta_j}\end{aligned}$$

- $||\xi|| = \langle \xi, \xi \rangle$ is a norm on the simplex.
- Aitchison distance $d_a(\xi, \eta) = \sqrt{\frac{1}{2k} \sum_i \sum_j \left(\log \frac{\xi_i}{\xi_j} - \log \frac{\eta_i}{\eta_j} \right)^2}$

  The inner product and norm are invariant to permutations of the components of the composition.

Log ratio transformations

- Additive log-ratio: $ALR(z) = \left(\log \frac{z_1}{z_k}, \ldots, log \frac{z_{k-1}}{z_k}\right)^T$
- Centered log-ratio: $CLR(z) = \left(\log \frac{z_1}{g(z)}, \ldots, log \frac{z_k}{g(z)}\right)^T$
- Isometric log-ratio transformation

Some models

- Measurement error model:

$$z_j = \xi \oplus \epsilon_j$$

- Regression:

$$\xi_j = \xi \oplus \gamma \otimes \mu_j$$

- Correspondence in Euclidean space:

$$\mu_j = \beta_0 + \beta_1(x_j - \bar{x})$$

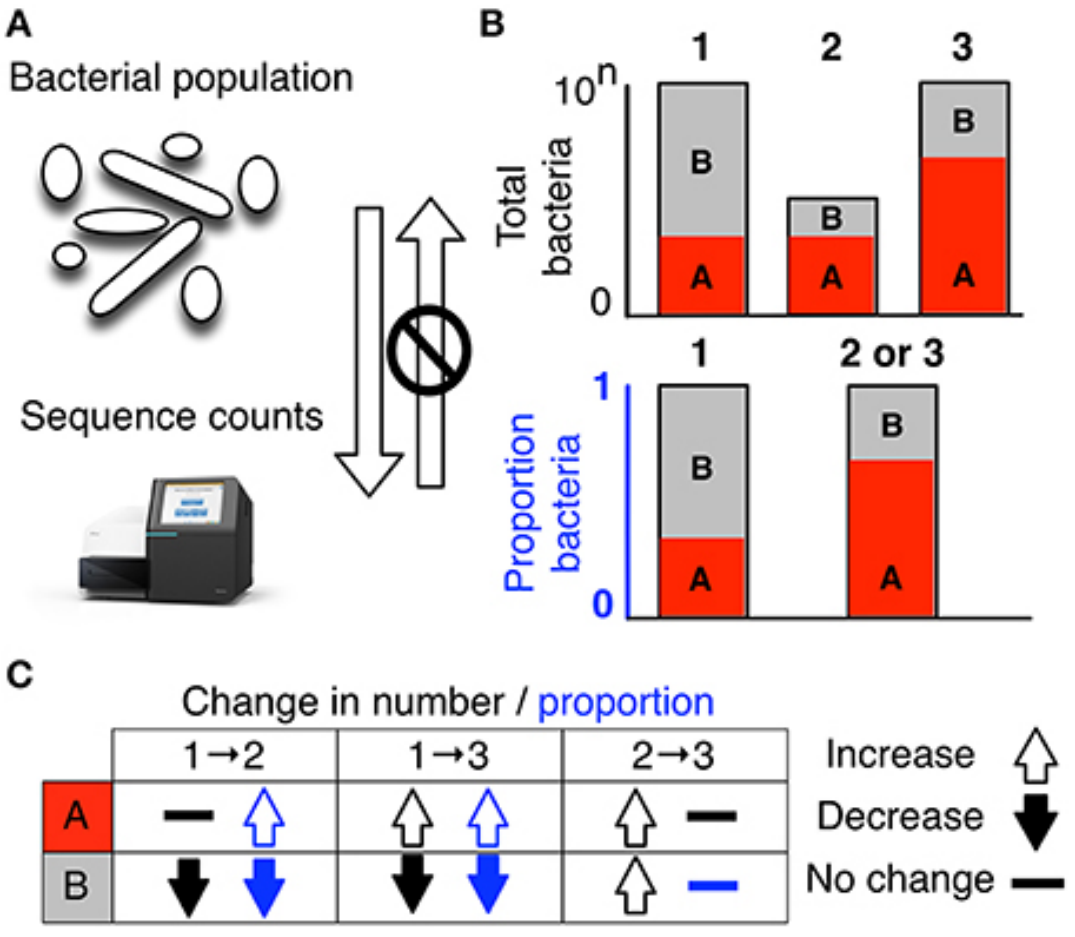
$$ALR^{-1}(\mu_j) = ALR^{-1}(\beta_0) \oplus ALR^{-1}(\beta_1) \otimes (x_j - \bar{x})$$

Distributions on a simplex

- If $ALR(z) = \left(\log \frac{z_1}{z_k}, \ldots, \log \frac{z_{k-1}}{z_k}\right)^T \sim MVN(\mu, \Sigma)$, z is logistic normal that $z \sim LN(\mu, \Sigma)$.
- Dirichlet distribution: an extension of the beta distribution. Ratios of independent Gammas
- Danish distribution: ratios of independent inverse Gaussian.
- Both with quite limited correlation structure.

Microbiome sequencing are compositional

- In ecological studies, it is possible that many species co-exist and their AA may be important.
- In high-throughput sequencing (HTS) experiments, sequencers deliver reads up to the capacity of the machine.
- Sequencing process can be considered as a random sampling of all molecule floating in the sample



Gloor, et al., 2017. *Frontier in Microbiology*

Compositional models in microbiome

Operation	Standard approach	Compositional approach
Normalization	Rarefaction 'DESeq'	CLR ILR ALR
Distance	Bray-Curtis UniFrac Jenson-Shannon	Aitchison
Ordination	PCoA (Abundance)	PCA (Variance)
Multivariate comparison	perManova ANOSIM	perMANOVA ANOSIM
Correlation	Pearson Spearman	SparCC SpiecEasi ϕ ρ
Differential abundance	metagenomSeq LEfSe DESeq	ALDEx2 ANCOM

Compositional models usually involve :

- Log-ratios (selecting one reference taxon, or CLR transformation)
- Pseudo-counts through simple or sophisticated imputations.

ANCOM

- Basic idea: if we form all pairwise log-ratios, for a taxon that is truly DA between conditions, many of the log-ratios with this taxon should be DA.
- For each taxon i, form all pairwise log-ratios with all other taxa.
- Assess the association between group membership with the log-ratios.
- Test statistics: number of log-ratios that is associated with clinical groups.

Linear log-contrast model

- Aitchison and Bacon-Shone (1984)
  Let Z represent the $n \times p$ microbiome data. Let $\mathbf{X}^p = \log(\frac{x_{ij}}{x_{ip}})$

$$y = \mathbf{X}^p \beta_{\setminus p} + \epsilon$$

  that $\beta_{\setminus p}$ is a $p-1$ dimensional coefficients.
- However, choosing the reference can be tricky.
- Alternative approach (Lin et al., 2014)

$$y = \mathbf{Z}\beta + \epsilon, \text{with} 1_p^T \beta = 0$$

Subcompositional regression models

- Let X_{gs} be the RA of s-th taxon within a higher rank g. $g = 1, \ldots, r, s = 1, \ldots, m_g$ such that

$$\sum_{s=1}^{m_g} X_{gs} = 1, \text{for } g = 1, \ldots, r$$

- X_g is the $n \times m_g$ matrix
-

$$Y = \sum_{g=1}^{r} Z_g \beta_g + \epsilon$$

that $Z_g = (Z_{g1}, \ldots, Z_{gm_g}) = (\log X_{g1}, \ldots, \log X_{gm_g}) \in R^{n \times m_g}$, $\beta_g = (\beta_{g1}, \ldots, \beta_{gm_g})^T$, we need the following constraints: $\mathbf{1}_{m_g}^T \beta_g = \sum_{s=1}^{m_g} \beta_{gs} = 0$ for $g = 1, \ldots, r$.

Penalized estimation

- General model

$$Y = \mathbf{Z}\beta + \epsilon, \text{subject to} \mathbf{C}^T\beta = 0$$

  In big n small p problem, solving constrained least squares is easy.

- Microbiome data is high dimensional –Sparse log-contrast regression

$$\hat{\beta}^n = \text{argmin}\left(\frac{1}{2n}||Y - \mathbf{Z}\beta||_2^2 + \lambda||\beta||_1\right), \quad \text{subject to} \quad \mathbf{C}^T\beta = 0$$

- Coordinate decent method of multipliers for optimization