Lit Review Notes:

Definitions/Revision:

Directional Data – Data on the boundary of hypershere i.e. unit vectors/directions

Correspondence Analysis: Dimensionality reduction for categorical variables, based on contingency tables.

- Row profiles are relative ratios of each column for a given row (analogous for column), average row/column profile (centroid) is the marginal row/column frequencies
- CA: represent chi square distance between individual row profiles and distance to average row profile (as with columns) graphically.
- "Inertia" or rows (identical for columns) is weighted sum of chi square distance of each row profile and average row profile (weighted by marginal frequency for that row) i.e. $\sum_i p_i d_i^2$, where d_i is chi squared distance from row i to row centroid
- Also equivalent to chi squared statistic/N
- CA: decompose into m dimensions, in decreasing order of their explained inertia/deviation from independence

Chi-Squared: Distribution of sum of squares of normal variables. Used to test differences between categorical variables (take squared difference of variable and expected value, which is assumed to be normal)

Paper Summaries:

Correspondence Analysis Biplots:

- Biplot: Low dimensional representation of rectangular matrix. Biplot is for rows and columns, not dimensionality of representation (though usually is in 2D). Points on a biplot (x_i,y_i) are used to reconstruct the ij-th entry of matrix by the scalar product x_i^Ty_i.
- Rank of matrix = dimensionality of perfect biplot reconstruction
- PCA and CA cases of biplot, loss based on correlation (variance) vs independence (Inertia)

Hypershpere Paper:

- Square root transformation maps compositional data to hypersphere
- Allows directional data distributions to model compositional
- Response (after transformation) given the covariates is modelled by Kent distribution

PCA for power transformed data:

Summary of techniques for CoDA:

- Aitchinson log transform
    - Interpretation of results exists relative to perspective change
    - Requires positive components
- Regression: Map to hypershere
- Representation Learning approach (more interpretable)
- Correspondence analysis

Example areas which are compositional:

- Genomics (see paper, NGS technologies lead to compositional data)
- Microbiome (microbiome paper also details this)
- Geological (e.g. rock/soil compositions)
- Economics

Research Communities:

http://www.compositionaldata.com/

https://www.coda-association.org/en/