Presentation for 2023 honours correspondence analysis section

Introduction to correspondence analysis (CA)

- Context: $\mathbf{Y}$ (counts)
  - Goal: Explain covariance structure of Y in terms of a smaller number of unobserved variables

Example datasets

- Suppose that we have some matrix $X : I \times J$ where each element is a count (or like it, in important ways)
  - Rows can be thought of as observations and columns as variables

Examples

- Rows are various dams, and columns are counts of waterbird species
- Rows are company brands (e.g. Cadbury, Beacon, Lindt), and columns are consumer ratings on a 1-5 scale (e.g. quality, price, taste)
- Rows are various immune compartments (e.g. blood, spleen, lymph), and columns are frequencies of immune cell types (e.g. T cells, B cells, NK cells)

Key characteristics

- Non-negative
- Natural zero
- Sparsity

Correspondence matrix, $\mathbf{P}$

- The correspondence matrix $\mathbf{P}$ is the matrix of overall proportions

$$\mathbf{P} = \frac{1}{\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}}\mathbf{X}$$

- Let $\mathbf{r}$ be the vector of row totals, i.e. $r_i = \sum_{j=1}^{n} P_{ij} = \mathbf{P1}$
- Let $\mathbf{c}$ be the vector of column totals, i.e. $c_j = \sum_{i=1}^{m} P_{ij} = \mathbf{P1}$

χ^2 distance

- The χ^2 distance is a measure for comparing two entities
- For example:
  - Comparing two histograms with equal bin placements
  - Comparing two densities
- In our case, we're comparing the observed counts (which are like a two-way histogram) and expected counts
- Formula:
  - $\sum_{ij} \frac{(x_{ij}-e_{ij})^2}{e_{ij}}$

CA vs PCA

- Association between rows and columns
  - PCA has that?
- Interested in profiles
  - PCA might standardise the variables, but not the rows as well
-

Examples of contexts for correspondence analysis