Correspondence analysis

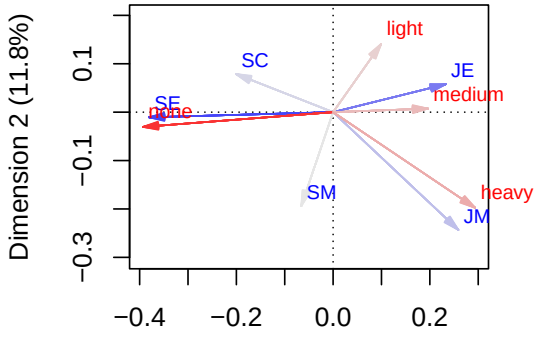
Introduction to correspondence analysis (CA)

- Context: $\mathbf{Y}$ (abundance)
  - Goal: Graphically display the relationships between and/or within the rows and columns

We go from this dataset:

	none	light	medium	heavy
SM	4	2	3	2
JM	4	3	7	4
SE	25	10	12	4
JE	18	24	33	13
SC	10	6	7	2

to this plot:



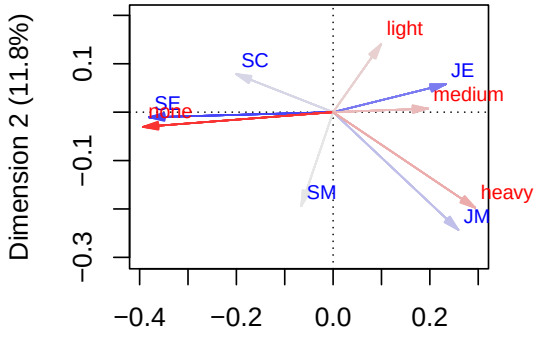
Introduction to correspondence analysis (CA)

- Context: $\mathbf{Y}$ (abundance)
  - Goal: Graphically display the relationships between and/or within the rows and columns

We go from this dataset:

	none	light	medium	heavy
SM	0.36	0.18	0.27	0.18
JM	0.22	0.17	0.39	0.22
SE	0.49	0.20	0.24	0.08
JE	0.20	0.27	0.38	0.15
SC	0.40	0.24	0.28	0.08

to this plot:



Example applications

Datasets

- Rows are various dams, and columns are counts of waterbird species
- 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)
- Rows are company brands (e.g. Cadbury, Beacon, Lindt), and columns are consumer ratings on a 1-5 scale (e.g. quality, price, taste)

Key characteristics

- Non-negative
- Natural zero (i.e. zero means literally nothing and not simply that two quantities are equal, for example)
- Same units (e.g. counts all in thousands)

The key property of the data is that proportions make sense throughout.

Correspondence matrix, $\mathbf{P}$

- Suppose that we have some matrix $\mathbf{X} : I \times J$ where each element
  - Rows can be thought of as observations and columns as variables
- The correspondence matrix $\mathbf{P} : I \times J$ is the matrix of overall proportions where

$$P_{ij} = \frac{x_{ij}}{\sum_{i=1}^{I} \sum_{j=1}^{J} x_{ij}} = \frac{x_{ij}}{n}$$

We go from $\mathbf{X}$

	none	light	medium	heavy
SM	4	2	3	2
JM	4	3	7	4
SE	25	10	12	4
JE	18	24	33	13
SC	10	6	7	2

to $\mathbf{P}$

	none	light	medium	heavy
SM	0.02	0.01	0.02	0.01
JM	0.02	0.02	0.04	0.02
SE	0.13	0.05	0.06	0.02
JE	0.09	0.12	0.17	0.07
SC	0.05	0.03	0.04	0.01

Independence of rows and columns

- Let $\mathbf{r}$ be the vector of row totals, i.e. $r_i = \sum_{j=1}^{J} P_{ij} = \mathbf{P1}$
- Let $\mathbf{c}$ be the vector of column totals, i.e. $c_j = \sum_{i=1}^{I} P_{ij} = \mathbf{P'1}$
- Then if the rows are independent of the cells, we have that

$$p_{ij} = r_i c_j, \implies \mathbf{P}_{\text{ind}} = \mathbf{rc'}$$

	none	light	medium	heavy
SM	0.02	0.01	0.02	0.01
JM	0.03	0.02	0.03	0.01
SE	0.08	0.06	0.08	0.03
JE	0.14	0.11	0.15	0.06
SC	0.04	0.03	0.04	0.02

Matrix of residuals

- Under the assumption of independence, we can calculate residuals:

$$\mathbf{P} - \mathbf{P}_{\text{ind}} = \mathbf{P} - \mathbf{r}\mathbf{c}'.$$

- Continuing the smoking example, we then have

$\mathbf{P}$

	none	light	medium	heavy
SM	0.02	0.01	0.02	0.01
JM	0.02	0.02	0.04	0.02
SE	0.13	0.05	0.06	0.02
JE	0.09	0.12	0.17	0.07
SC	0.05	0.03	0.04	0.01

$\mathbf{P} - \mathbf{r}\mathbf{c}'$

	none	light	medium	heavy
SM	0.003	-0.003	-0.003	0.003
JM	-0.009	-0.006	0.006	0.009
SE	0.046	-0.010	-0.023	-0.014
JE	-0.051	0.018	0.025	0.008
SC	0.011	0.001	-0.005	-0.006

- Residuals are naturally larger for the more abundant rows (employee ranks)