⑤ Distance

Many types of distance

Other distance metrics

Mahalanobis (or statistical) distance

$$\sqrt{((X_A - X_B)^T S^{-1}(X_A - X_B))}$$

Manhattan:

$$\sum_{j=1}^{p} |(X_{aj} - X_{bj})|$$

Minkowski:

$$(\sum_{j=1}^{p} |(X_{aj} - X_{bj})|^m)^{1/m}$$

Distances for count data

Canberra:

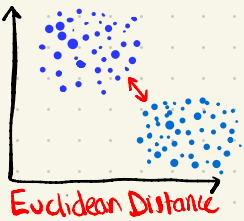
$$\frac{1}{n_z} \sum_{j=1}^{p} \frac{X_{aj} - X_{bj}}{X_{aj} + X_{bj}}$$

Bray-Curtis:

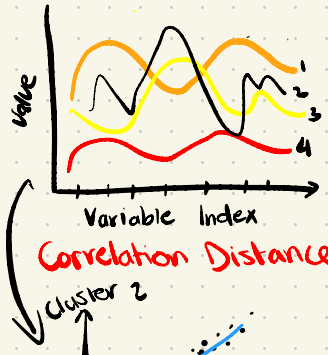
$$\frac{\sum_{j=1}^{p} |X_{aj} - X_{bj}|}{\sum_{j=1}^{p} (X_{aj} + X_{bj})}$$

- This slide might be in next weeks lecture so don't stress this too much.
- Math is not particularly important.

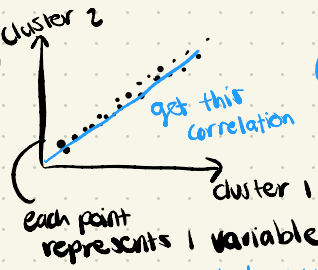
Only (really) need to know ①



as a scatter plot



Correlation Distance



each point represents 1 variable

(counter intuitive to how we think of scatter plots & correlation)

Correlation dist:

$d = 1 - r_{C_1C_2}$ ← correlation ranges −1 to 1

$d = 0$ when $r_{C_1C_2} = 1$ (perfect correlation)

$d = 2$ when $r_{C_1C_2} = -1$ (perfect negative correlation)