Non-metric multidimensional scaling

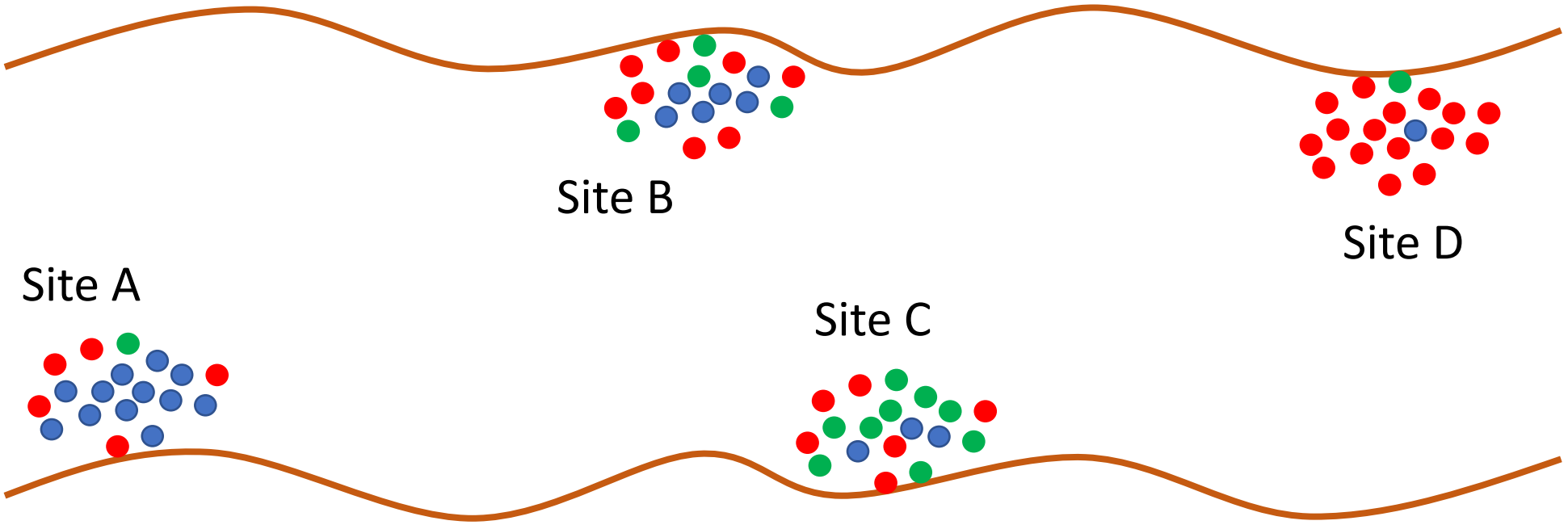
Problem Statement

Many ordination methods are available to reduce dimensionality in datasets. Multidimensional scaling is used to examine **similarity** between data points

A key problem facing biologists is how to analyse community data, where abundance of species/genes/microbes is measures at different sites/patients/samples

Such data are difficult to analyse as communities are often composed of many different species/genes/microbes

Problem Statement



Communities of bacteria in the gut: how to compare relative abundance?

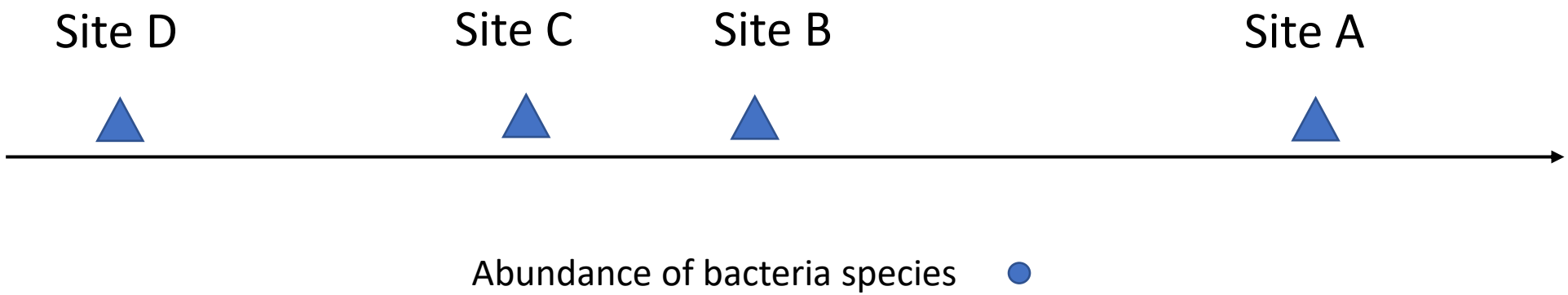
NMDS and high dimensional data

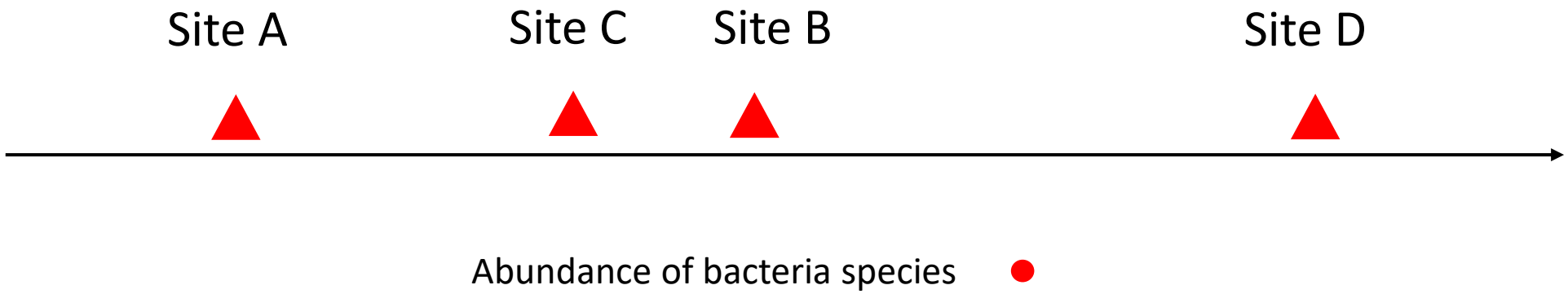
MDS is a useful method to calculate amount of similarity (or dissimilarity) in community composition between sites/patients/samples by calculating distance between pairs of points

Multidimensional scaling uses actual distances between points, while non-metric multidimensional scaling (NMDS) uses ranks of distances

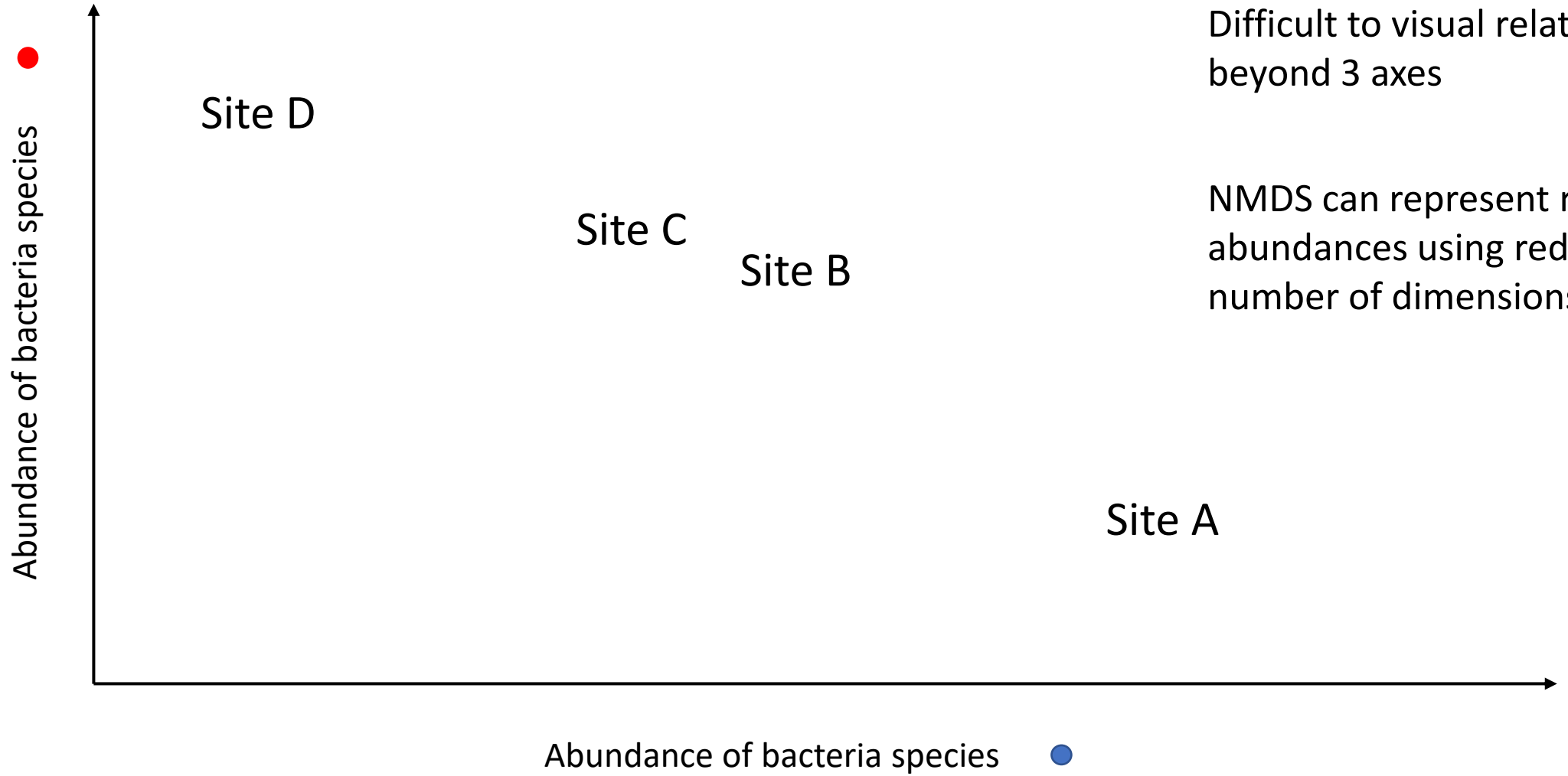
NMDS is flexible and makes very few assumptions about the data (can use continuous and categorical variables, does not assume linear relationships, can include missing data)

NMDS and high dimensional data





NMDS and high dimensional data



Difficult to visual relationships beyond 3 axes

NMDS can represent relative abundances using reduced number of dimensions/axes

Carrying out NMDS

1) Map relative positions of communities/sites in all dimensions

2) Choose reduced number of dimensions to represent data

3) Calculate difference between ranks on all dimensions vs ranks on reduced dimensions

4) Compare ranks on all dimensions vs ranks on reduced dimensions

5) Calculate amount of disagreement between ranks using all dimensions and ranks using reduced number of dimensions (stress)

Example

We will use the 'microbiome' package from Bioconductor to carry out a NMDS analysis. The microbiome R package allows us to explore and analyse some microbiome profiling data collected from African Americans and rural Africans (dietswap dataset).

Contains data from **222** samples taken from 222 participants containing abundance data from **130** different taxa.

The data contains phylogenetic sequencing data, as well as metadata from study participants

We select the most prevalent taxa in the samples for analysis

Challenge 1

What do you think we could do to reduce stress if the overall solution of the ordinate function had a stress solution greater than 0.3?

Discuss in groups

Challenge 2

Identify the lowest value of k that gives the best stress value using the microbiome data (ds.core). Think about getting a low stress value as well as creating results that can be easily visualised and interpreted.

Use the function stressplot to examine how distribution around the regression line changes with increases in k. Use what you have learned to decide on best number of dimensions to include in NMDS analysis of the microbiome data. Use ‘set.seed(4000)’ for k values >2.

Challenge 3

Use the 'plot_ordination' function to look for clusters in points on the NMDS axes according to sex, bmi_group and other factors in the metadata. Do any of these factors appear to form clusters?

Challenge 4

What conclusions can you draw from this NMDS analysis? Does there appear to be a difference in the microbiome communities of rural Africans and African Americans? What other factors may be important determinants in diversity of microbiome communities?

Further reading

Holland, S.M. (2008) Non-metric Multidimensional Scaling (MDS). https://strata.uga.edu/software/pdf/mdsTutorial.pdf

Lahti, L, Sudarshan, S. et al. Introduction to the microbiome R package. https://microbiome.github.io/tutorials/

Marco-Ramell, A. et al. (2018) Evaluation and comparison of bioinformatic tools for the enrichment analysis of metabolomics data. BMC Bioinformatics 19:1. DOI:10.1186/s12859-017-2006-0

O'Keefe, S.J.D. et al. (2015) Fat, fibre and cancer risk in African Americans and rural Africans. Nature Communications 6:6342. DOI:10.1038/ncomms7342

Oksanen, J. (2005) Multivariate analysis of Ecological Communities in R. http://www.pelagicos.net/MARS6910_spring2015/manuals/R_vegan_multivariate.pdf