Multi-dimensional scaling (MDS)

Pablo Aguilar

Pablo.aguilar@uantof.cl
Department of Biotechnology
University of Antofagasta

Multidimensional scaling (MDS) is a set of data analysis techniques used to explore the structure of (dis)similarity data.

MDS represents a set of objects as points in a multidimensional space in such a way that the points corresponding to similar objects are located close together, while those corresponding to dissimilar objects are located far apart.

Goal of Multidimensional scaling (MDS): Given pairwise dissimilarities, reconstruct a map that preserves distances.

Multidimensional scaling

- MDS is a family of different algorithms, each designed to arrive at optimal low-dimensional configuration (p = 2 or 3)

- MDS methods include

1 Classical MDS

2 Metric MDS

3 Non-metric MDS

Distance, dissimilarity and similarity

Distance, dissimilarity and similarity (or proximity) are defined for any pair of objects in any space. In mathematics, a distance function (that gives a distance between two objects) is also called metric, satisfying

- $d(x, y) \geq 0$
- $d(x, y) = 0$ if and only if $x = y$
- $d(x, y) = d(y, x)$
- $d(x, z) \leq d(x, y) + d(y, z)$

Given a set of dissimilarities, one can ask whether these values are distances and, moreover, whether they can even be interpreted as Euclidean distances (i.e., simply the straight-line distance between two points in multivariate space)

Euclidean and non-Euclidean distance

Given a dissimilarity (distance) matrix $D = (d_{ij})$, MDS seeks to find $x_1, \ldots, x_n \in R^p$ so that

$$dij \approx \|xi - xj\|_2 \text{ as close as possible.}$$

Oftentimes, for some large p, there exists a configuration $x_1, \ldots, x_n$ with exact distance match $dij \equiv \|xi - xj\|_2$. In such a case the distance d involved is called a Euclidean distance.

There are, however, cases where the dissimilarity is distance, but there exists no configuration in any p with perfect match

$$d_{ij} \not= \|x_i - x_j\|_2 \text{ , for some i , j.}$$

Such a distance is called non-Euclidean distance.

Nevertheless, MDS seeks to find an optimal configuration x_i that gives $d_{ij} \approx \|x_i - x_j\|_2$ as close as possible.

Nonmetric multidimensional scaling (nMDS)

Non-metric: Without axes.

Multidimensional: Represents relationships between multiple variables in two or three dimensions.

Scaling: The ratio between reality and representation.

How does nMDS work?

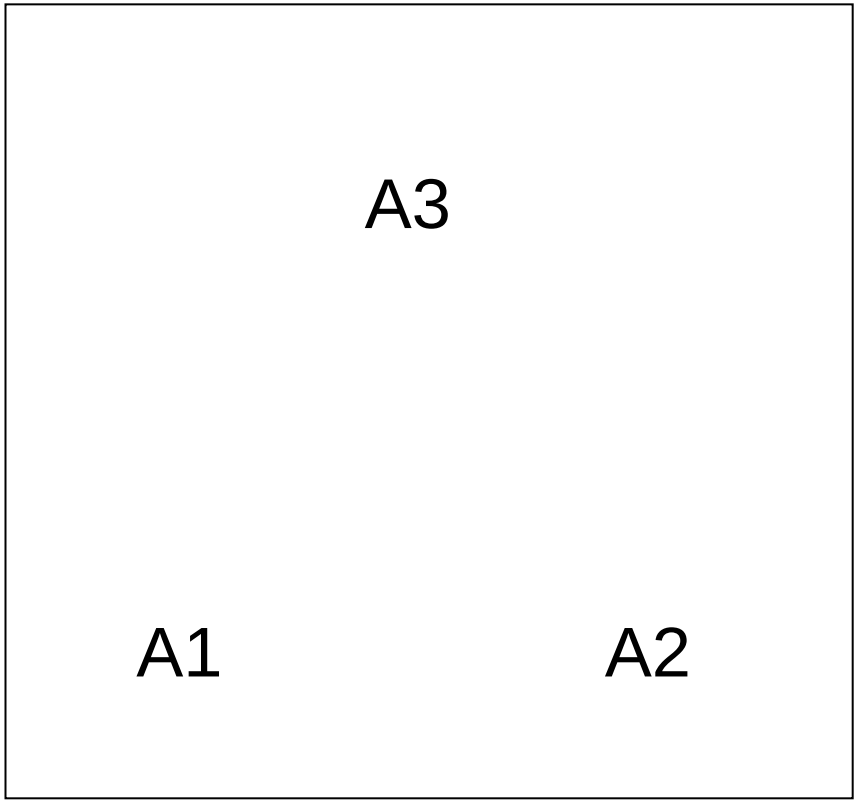
nMDS uses the RANK ORDER of similarity relationships between samples.

Sample	Sample	% sim	rank
A1	A2	99%	1
A1	A3	96%	2
A2	A3	95%	3

A1 closer to A2 than A3

How does nMDS work?

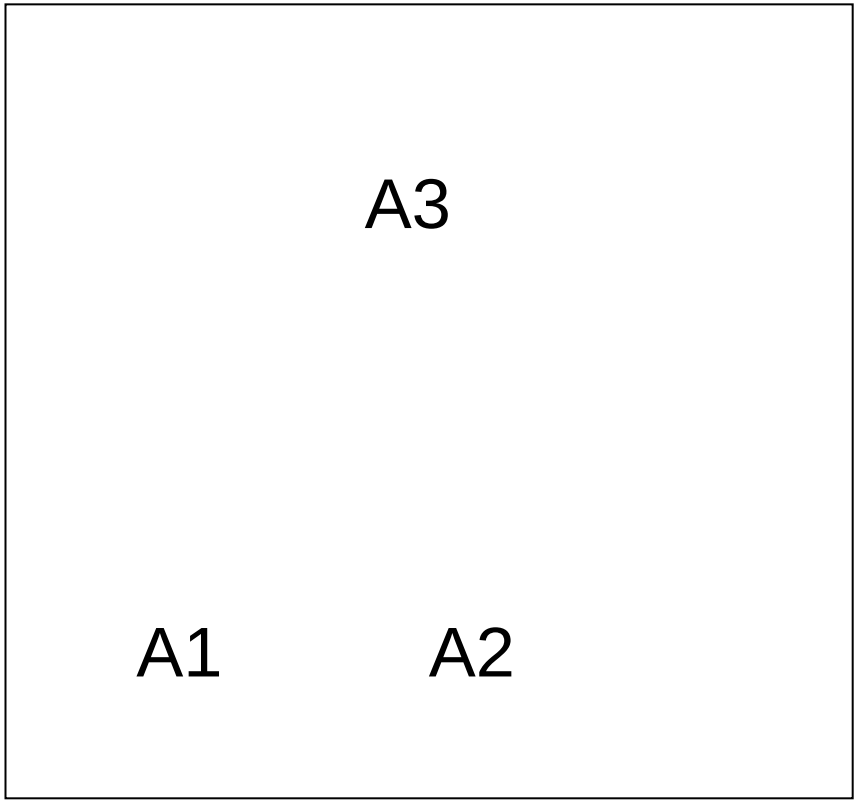
Next, nMDS places the points in a 2 (or 3)-dimensional space to represent this rank order.



A1 closer to A2 than A3

How does nMDS work?

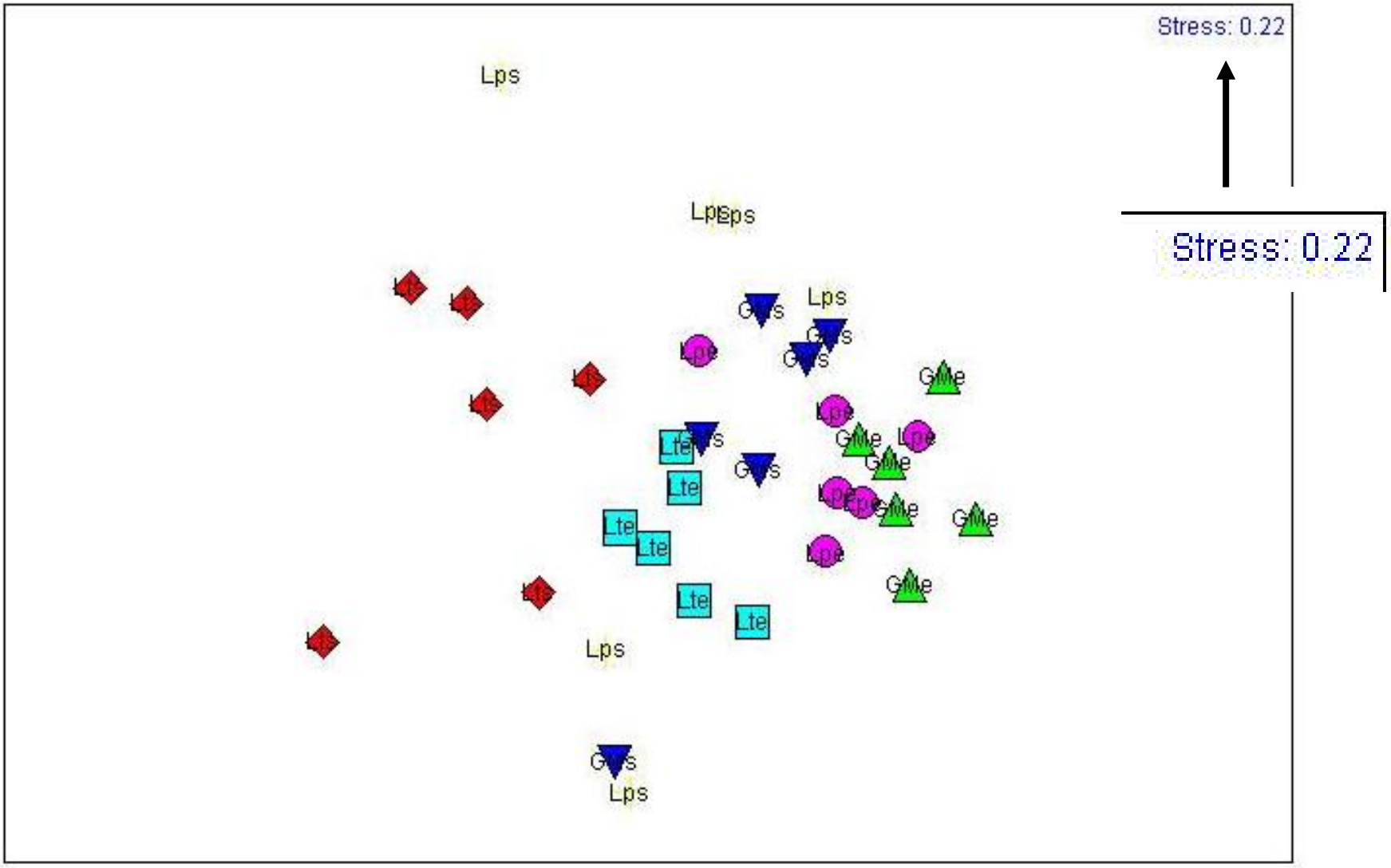
Next, nMDS places the points in a 2 (or 3)-dimensional space to represent this rank order.



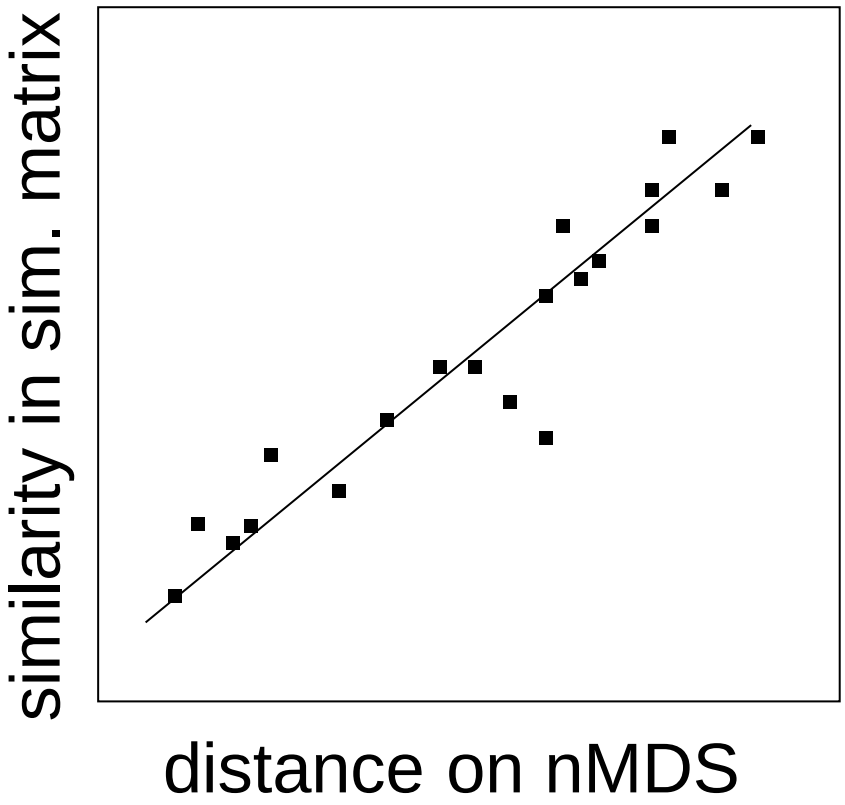
A1 closer to A2 than A3

How accurate is the nMDS map?

Sometimes nMDS cannot represent all relationships accurately.



This is reflected in a high STRESS value.



If the stress value is:

0.0: Perfect map

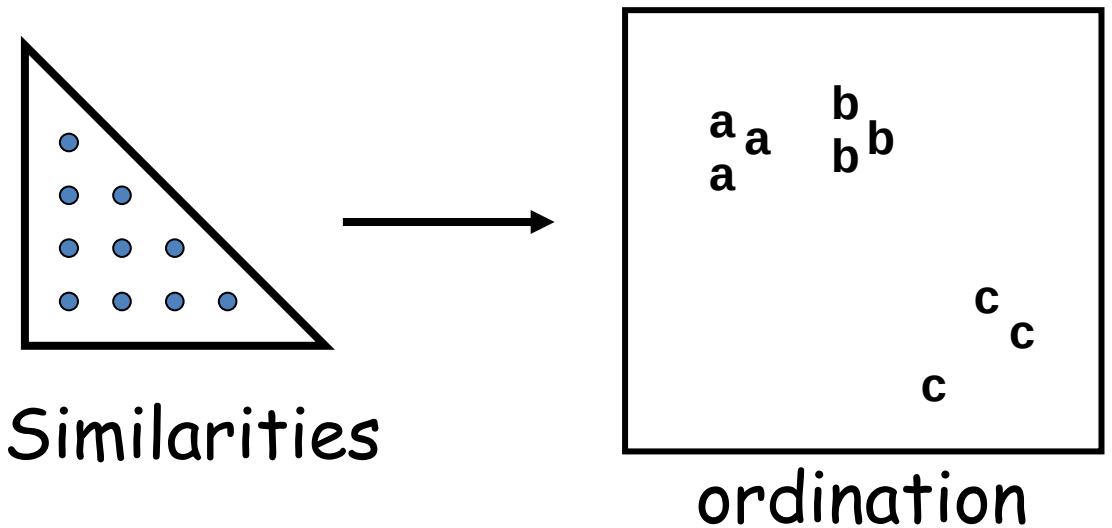
0.1: Decent map

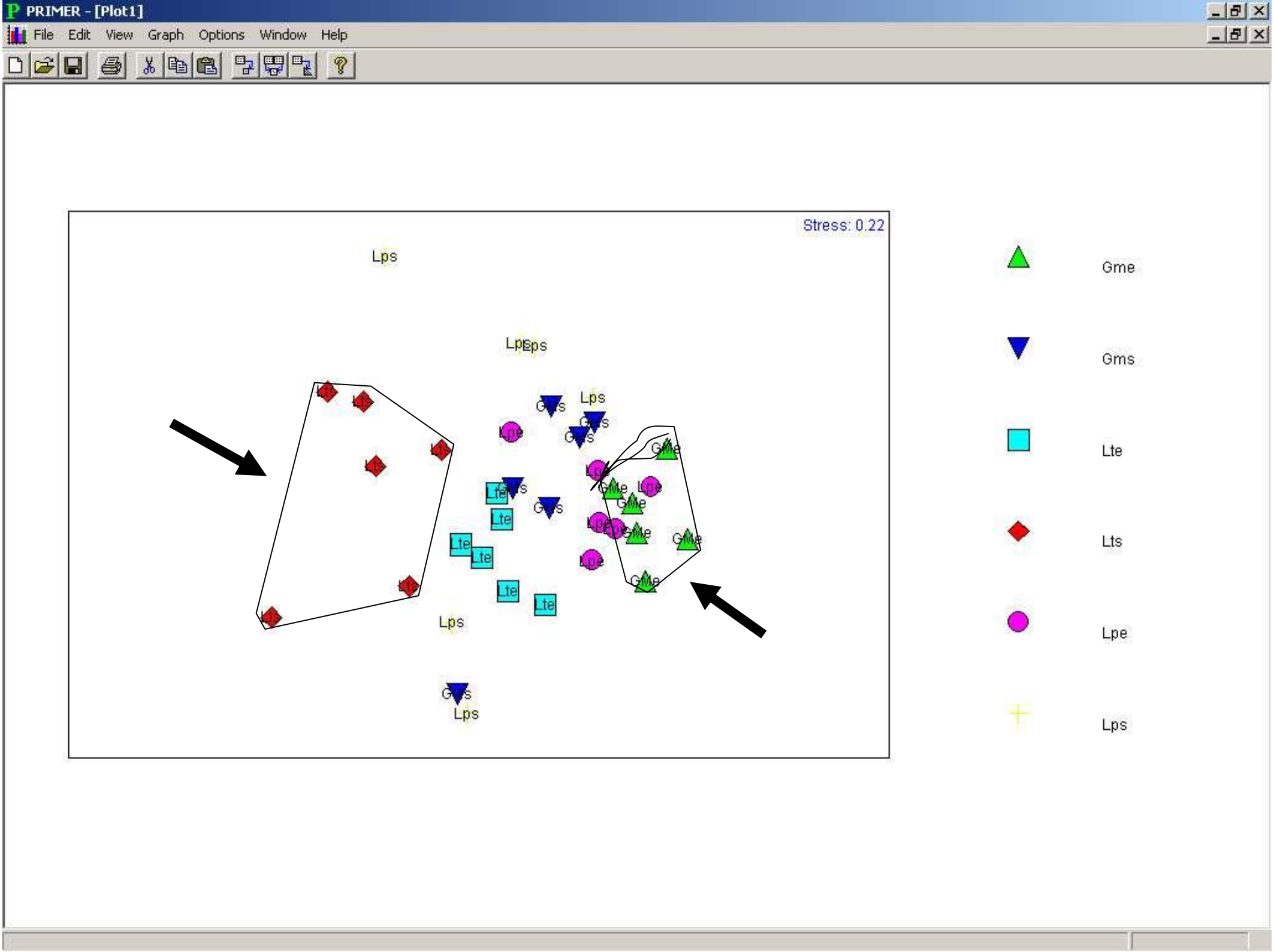
0.2: Good map

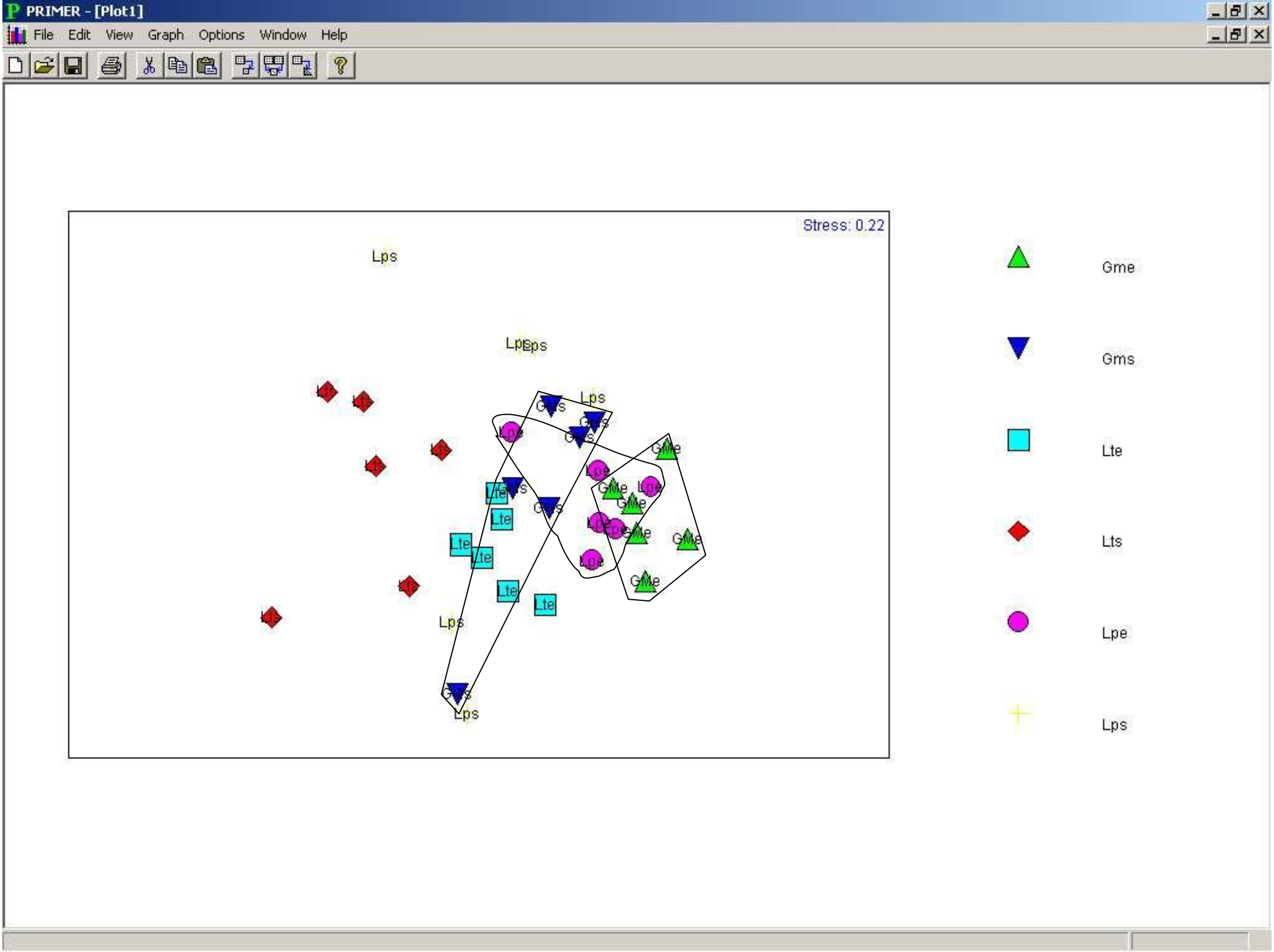
0.3: Not good

Important considerations about Ordination analysis

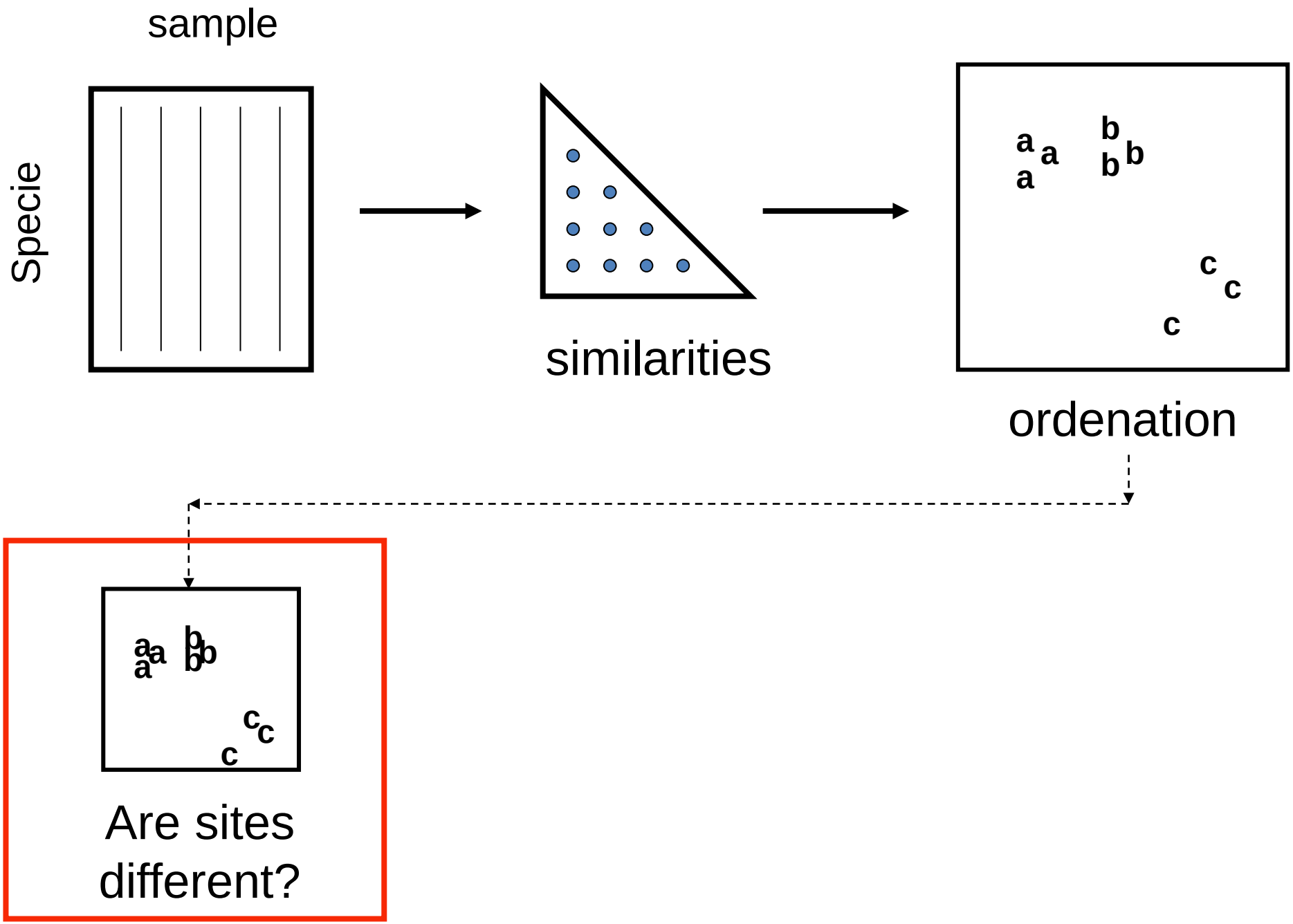
- Ordination is a way to visualize the similarity between our samples
- nMDS aims to visually represent the rank order obtained within the similarity matrix.
- What matters is the distance between the points.
- The stress value allows estimating the 'quality' of the nMDS.




PRIMER - [Plot1]
File Edit View Graph Options Window Help
Stress: 0.22
Lps
Lps
Lps
Lps
Lps
Gme
Gms
Lte
Lts
Lpe
Lps


PRIMER - [Plot1]
File Edit View Graph Options Window Help
Stress: 0.22
Lps
Lps
Lps
Lps
Lps
Lps
Gme
Gms
Lte
Lts
Lpe
Lps

Secuencia del Analisis



Time for nMDS exercises !!

https://github.com/PabloPablo-Aguilar/nMDS

1. dataset: Herbivore

Abundance of herbivore in different habitats.

2. dataset: lakes

Microbial community composition of lakes located at different geographic zones

3. dataset: ALPS_annual / ALPS_season

The dataset contains observations collected from an alpine lake over several months. Each observation the microbial species abundances.

For the purpose of the assignment, the focus is on comparing samples collected during two distinct periods: the ice cover season and the free ice season of the lake. The ice cover season refers to the period when the lake surface is frozen, typically during winter months, while the free ice season refers to the period when the lake is not frozen, typically during spring, summer, and autumn months.

Are the microbial communities different between both season (ice cover/ free)?