Outline

Multidimensional Scaling (MDS)

MDS is a class of methods based on representing high-dimensional data in a lower dimensional space so that inter-point distances are preserved as "best" as possible. MDS effectively "squeezes" a high-dimensional cloud of points into a smaller number of dimensions, generally 2 or 3.
Given $x_1, \ldots, x_n \in \mathbb{R}^p$, we can obtain a matrix of pairwise distances D with entries $d_{ij} = d(x_i, x_j)$ using some measure of dissimilarity d. For example Euclidean distance $d_{ij} = \|x_i - x_j\|_2$. In most applications, only D is available. MDS finds representations $z_1, \ldots, z_n \in \mathbb{R}^k$ such that

$$d(x_i, x_j) \approx \tilde{d}(z_i, z_j),$$

where d represents dissimilarity in the original p-dimensional space and $\tilde{d}$ represents dissimilarity in the reduced k-dimensional space. The 'best' values of z_i are chosen to minimise some *stress function*.

Metric vs Non-Metric Stress Functions

Metric

Where closeness is considered geometrically, Euclidean distance $d_{ij} = \|x_i - x_j\|_2$ is commonly measured with the classical stress function

$$S_{\text{metric}}(d_{ij}, \tilde{d}_{ij}) = \sum_{i \neq j} \left(d_{ij} - \tilde{d}_{ij}\right)^2$$

Non-Metric

Sometimes it is more important to retain the ordering of d_{ij} as good as possible rather than the actual values assigned. Non-metric stress functions have been developed for ordered distances

$S_{\text{non-metric}}(d_{ij}, \tilde{d}_{ij}) = \min_{g \text{ monotone}} \frac{1}{\sum_{i \neq j} \tilde{d}_{ij}^2} \sum_{i \neq j} \left(g(d_{ij}) - \tilde{d}_{ij}\right)^2$

Solving the Metric MDS Problem

Suppose we only have an $n \times n$ matrix of Euclidean distances $D = (d_{ij})$ but not the points X themselves. The Classical MDS problem is to find a configuration of n points in p-dimensional space that yields the same Euclidean distance matrix as X.

Infinitely many solutions exist as the distance matrix is invariant to rigid motions (rotations, reflections and translations).

As distances are Euclidean, can write $d_{ij} = \|x_i - x_j\|_2$ for some points $x_1, \ldots, x_n \in \mathbb{R}^p$, where

$$\begin{aligned} d_{ij}^2 &= \|x_i - x_j\|_2^2 \\ &= (x_i - x_j)(x_i - x_j)^\top \\ &= x_i x_i^\top + x_j x_j^\top - 2 x_i x_j^\top \end{aligned} \tag{1}$$

Solving the Metric MDS Problem

We define matrix B with entries $b_{ij} = x_i x_j^\top$, we can compute D from B but also B from D. From this, it is possible to recover a configuration which solves this problem.
Writing (1) in terms of b_{ij}, we have

$$d_{ij}^2 = b_{ii} + b_{jj} - 2b_{ij} \tag{2}$$

$\Rightarrow$ If two configurations of n objects in p-dimensional space have identical matrix $B = XX^\top$, then they also share the same distance matrix D.

We can also compute b_{ij} in terms of d_{ij} assuming $\sum_i x_i = 0$ (problem sheet).

Solving the Metric MDS Problem

If two configurations of n objects in p-dimensional space have identical matrix $B = XX^\top$, then they also share the same distance matrix D.
Considering the eigendecomposition of B, we see that $B = XX^\top = ULU^\top$ for some orthogonal matrix U with columns $U = (u_1, \ldots, u_n)$ and diagonal matrix L with entries $\lambda_1, \ldots, \lambda_n$. [1]
So if $n > p$ we can write

$$\tilde{X} = [\sqrt{\lambda_1}U_1, \ldots, \sqrt{\lambda_p}U_p]$$

i.e. we have found a p-dimensional configuration of n points $\tilde{X}$ with the *same* distance matrix D as X.

[1] If $X = UDV^\top$ is again the SVD of X, then $XX^\top = UDD^\top U^\top$. The matrix U is thus the same in the EVD of $XX^\top$ and the $n \times n$-matrix $L = DD^\top$ has the same diagonal entries as the $p \times p$-matrix $\Lambda = D^\top D$ in the SVD of $X^\top X$.

MDS Example: US City Flight Distances

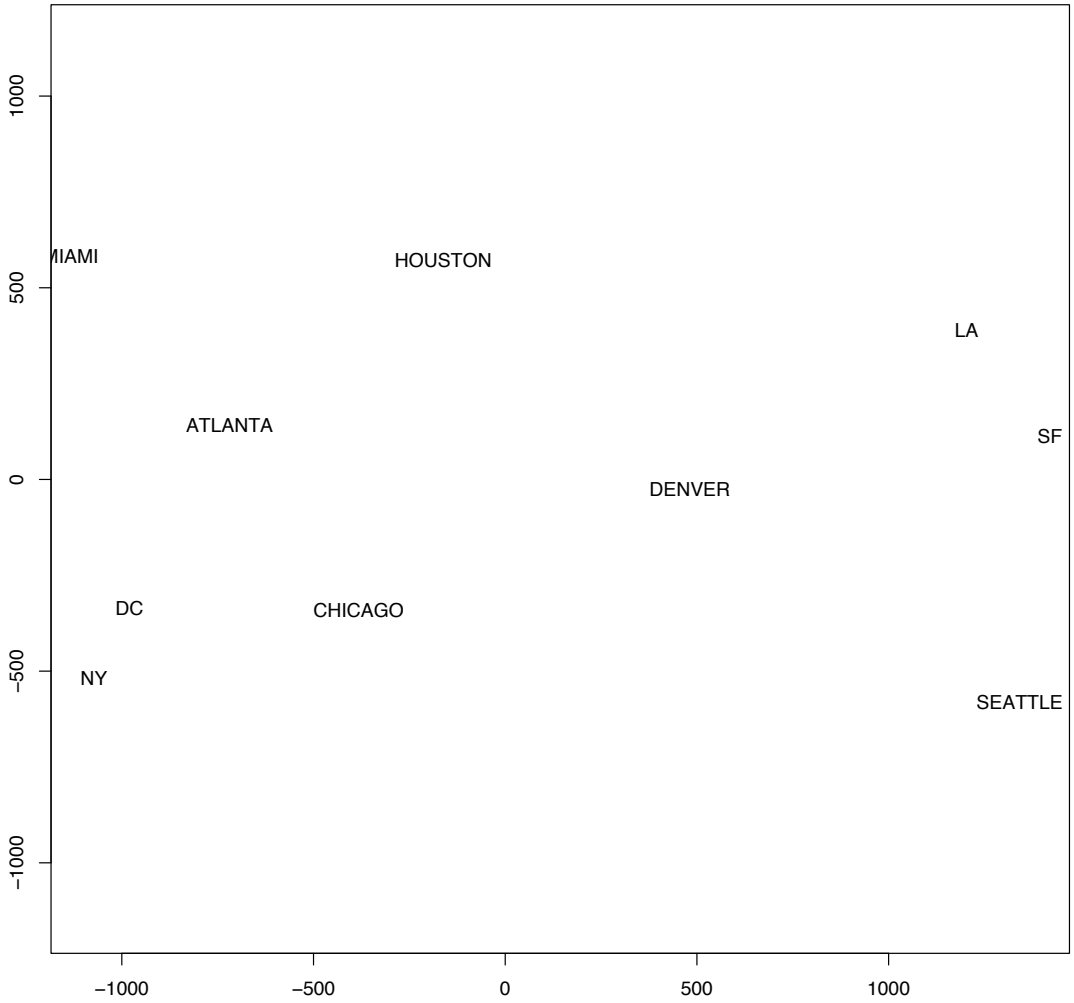
We present a table of flying mileages between 10 American cities, distances calculated from our 2-dimensional world. Using D as the starting point, metric MDS finds a configuration with the same distance matrix.

ATLA	CHIG	DENV	HOUS	LA	MIAM	NY	SF	SEAT	DC
0	587	1212	701	1936	604	748	2139	2182	543
587	0	920	940	1745	1188	713	1858	1737	597
1212	920	0	879	831	1726	1631	949	1021	1494
701	940	879	0	1374	968	1420	1645	1891	1220
1936	1745	831	1374	0	2339	2451	347	959	2300
604	1188	1726	968	2339	0	1092	2594	2734	923
748	713	1631	1420	2451	1092	0	2571	2408	205
2139	1858	949	1645	347	2594	2571	0	678	2442
2182	1737	1021	1891	959	2734	2408	678	0	2329
543	597	1494	1220	2300	923	205	2442	2329	0

MDS Example: US City Flight Distances

```
library(MASS)

us <- read.csv("http://www.stats.ox.ac.uk/
          ~teh/teaching/datamining/data/uscities.csv")

## use the classical stress function
## to find lower dimensional views of the data
## recover X in 2 dimensions

us.classical <- cmdscale(d=us,k=2)

plot(us.classical)
text(us.classical,labels=names(us))
```

MDS Example: US City Flight Distances



Lower-dimensional Reconstructions

Having managed to reconstruct a set of p-dimensional points with the same distance matrix D, we would like to find lower dimensional representations which minimise the stress function S_{metric}.
If the SVD of X is given by $X = UDV^\top$, then

$$B = XX^\top = UDD^\top U = ULU^\top$$

Generally the representation of $\tilde{X}$ (chosen so that $\tilde{X}$ and X have the same distance matrix) can be written as

$$\tilde{X} = [\sqrt{\lambda_1}U_1, \ldots, \sqrt{\lambda_r}U_r]$$

where r is the rank of B.
Setting the smallest eigenvalues to zero reveals the 'best' k-dimensional view of the data (where k is the number of non-zero eigenvalues), minimizing the stress function (proof not given).
This is analogous to PCA, where the smallest eigenvalues of $X^\top X$ are effectively suppressed. Indeed, both PCA and MDS under Euclidean distance are dual and yield effectively the same result (yet MDS can also be applied to distance matrices not generated under Euclidean distance measure).

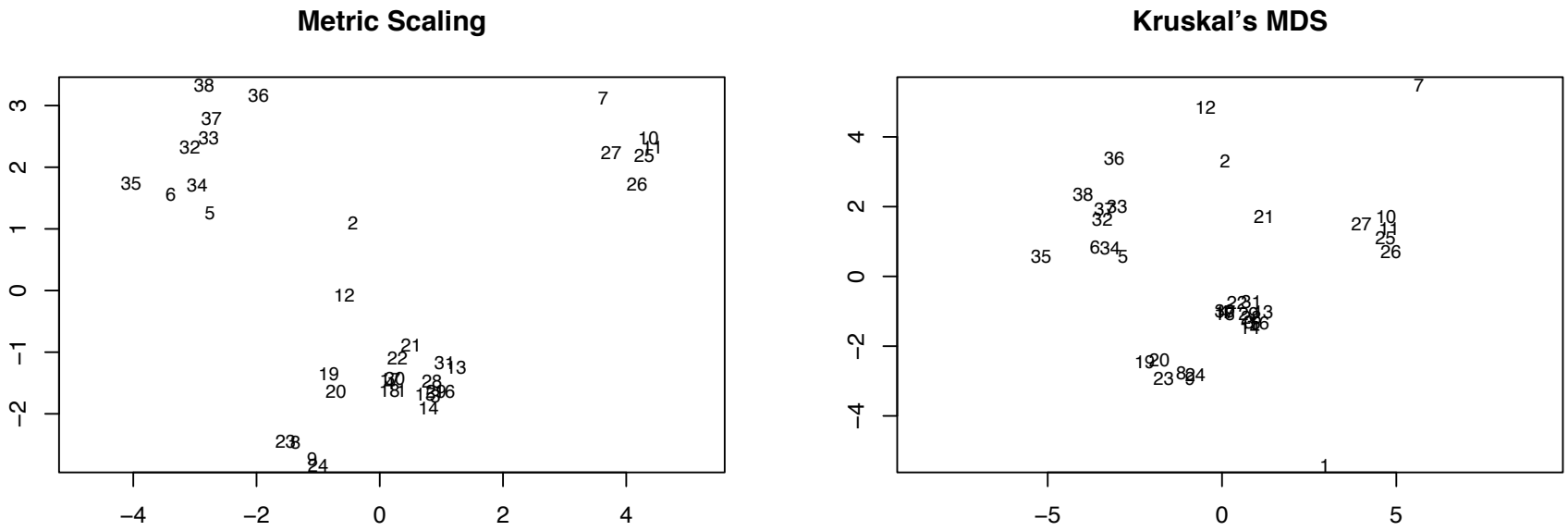
MDS Example: Virus Data

A data set on 39 viruses with rod-shaped particles affecting various crops (tobacco, tomato, cucumber and others), described by Fauquet *et al.* (1988). These are *Tobamoviruses* with monopartite genomes spread by contact.

There are 18 measurements on each virus, the number of amino acid residues per molecule of coat protein; the data come from a total of 26 sources.

We want to investigate whether there are subgroups within this group of viruses.

MDS Example: Virus Data



Distance-based representations of the *Tobamovirus* group of viruses (the variables were scaled before Euclidean distance was used).

MDS Example: Virus Data

MDS reveals some clear subgroups within the *Tobamoviruses*.

Viruses 7 (cucumber green mottle mosaic virus) and 21 (pepper mild mottle virus) have been clearly separated from the other viruses in the non-metric MDS plot, which is not the case in the metric version.

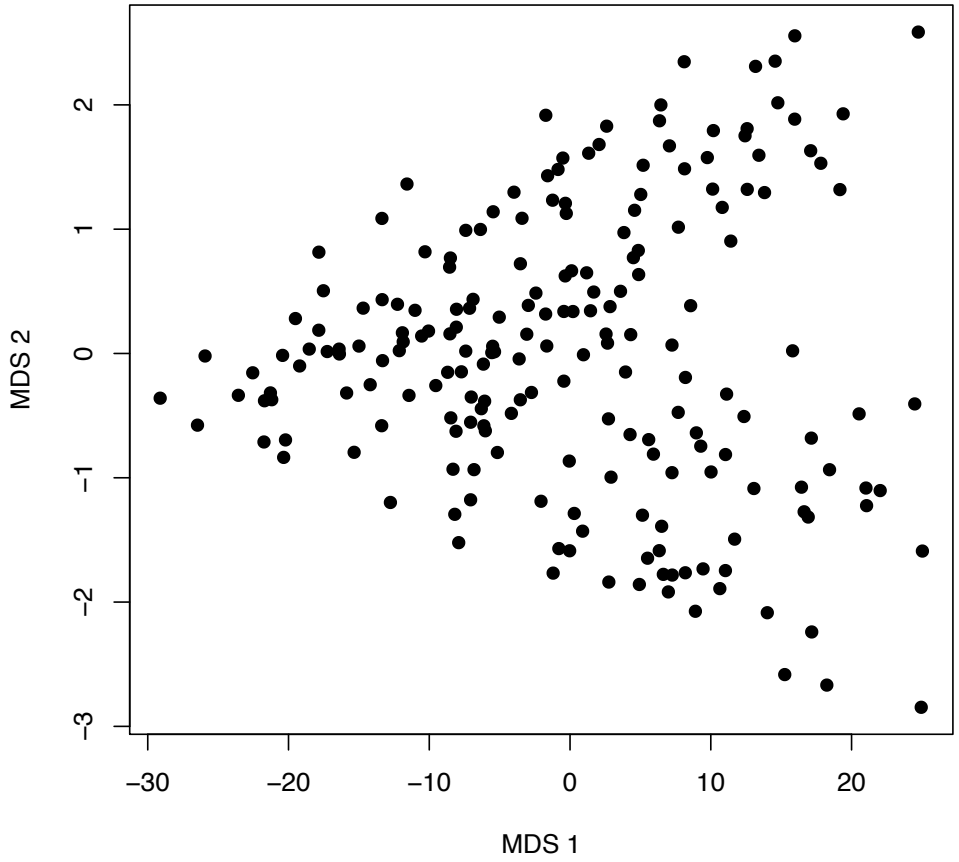
Ripley (1996) states that the non-metric MDS plot shows interpretable groupings. The upper right is the cucumber green mottle virus, the upper left is the ribgrass mosaic virus. The one group of viruses at the bottom, namely 8,9,19,20,23,24, are the tobacco mild green mosaic and odontoglossum ringspot viruses.

Example: Crabs data

```
library(MASS)
Crabs <- crabs[,4:8]
Crabs.class <- factor(paste(crabs[,1],crabs[,2],sep=""))

crabsmds <- cmdscale(d= dist(Crabs),k=2)
plot(crabsmds, pch=20, cex=2)
```

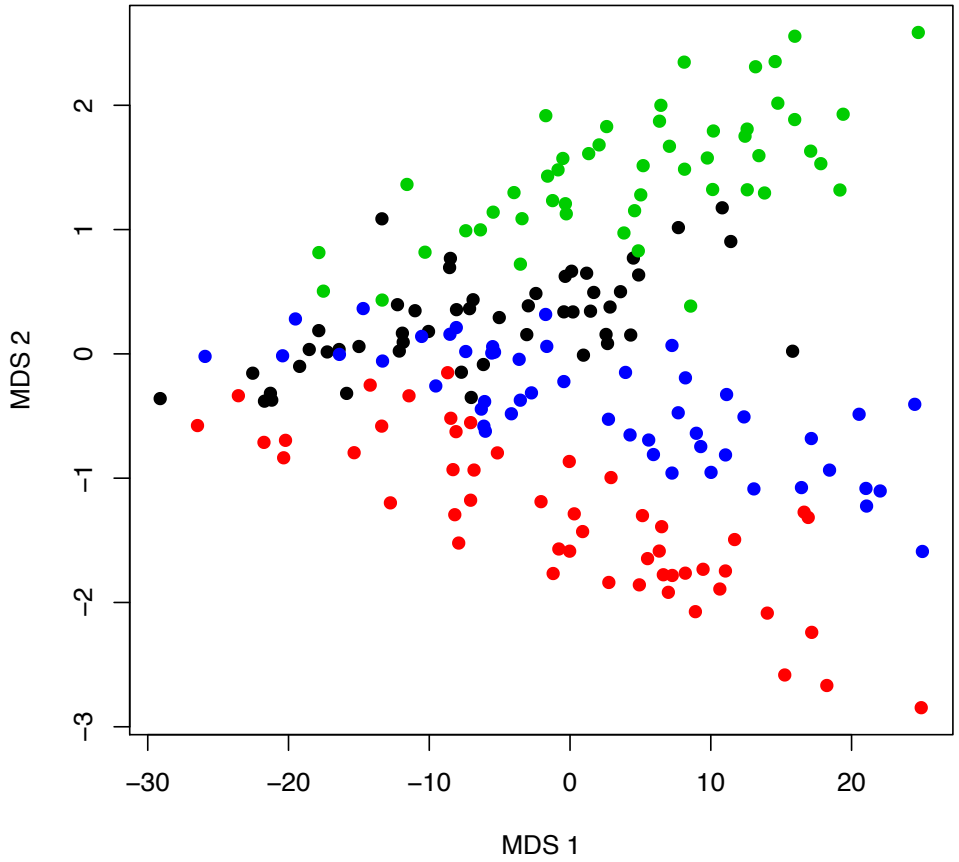
Example: Crabs data

First two MDS components.



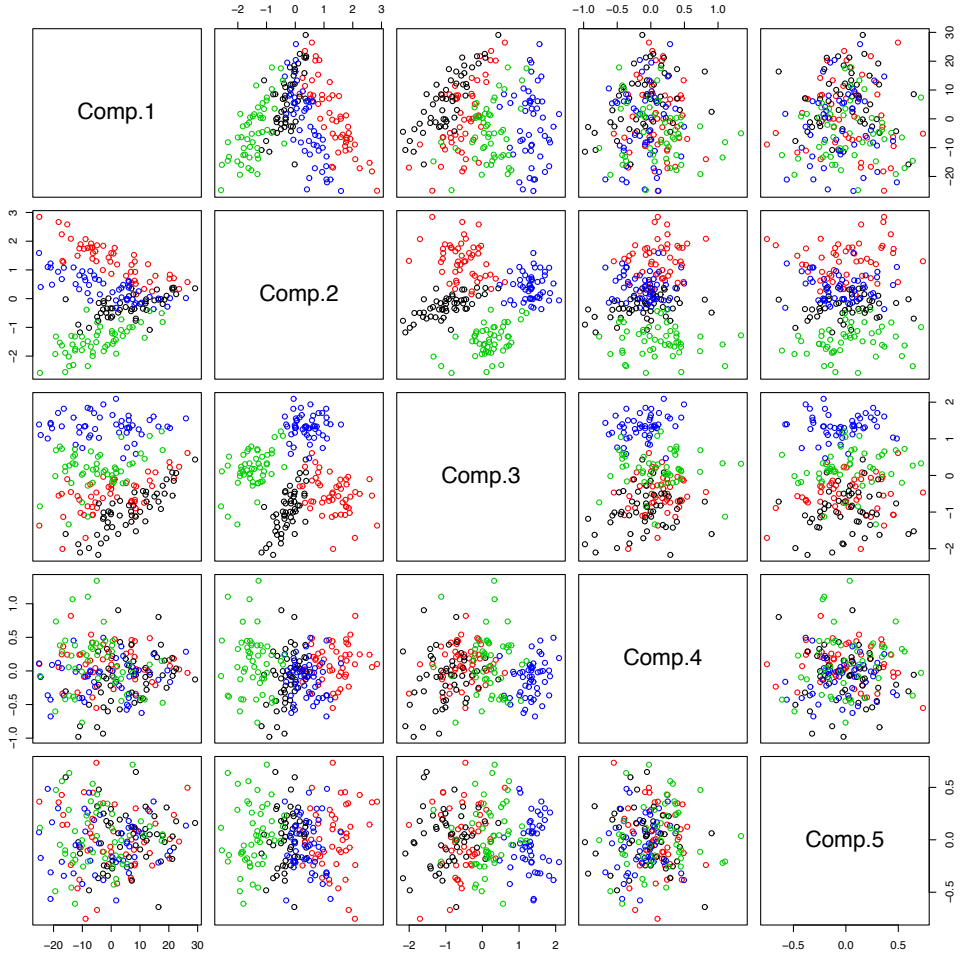
Example: Crabs data

With grouping information.



Example: Crabs data

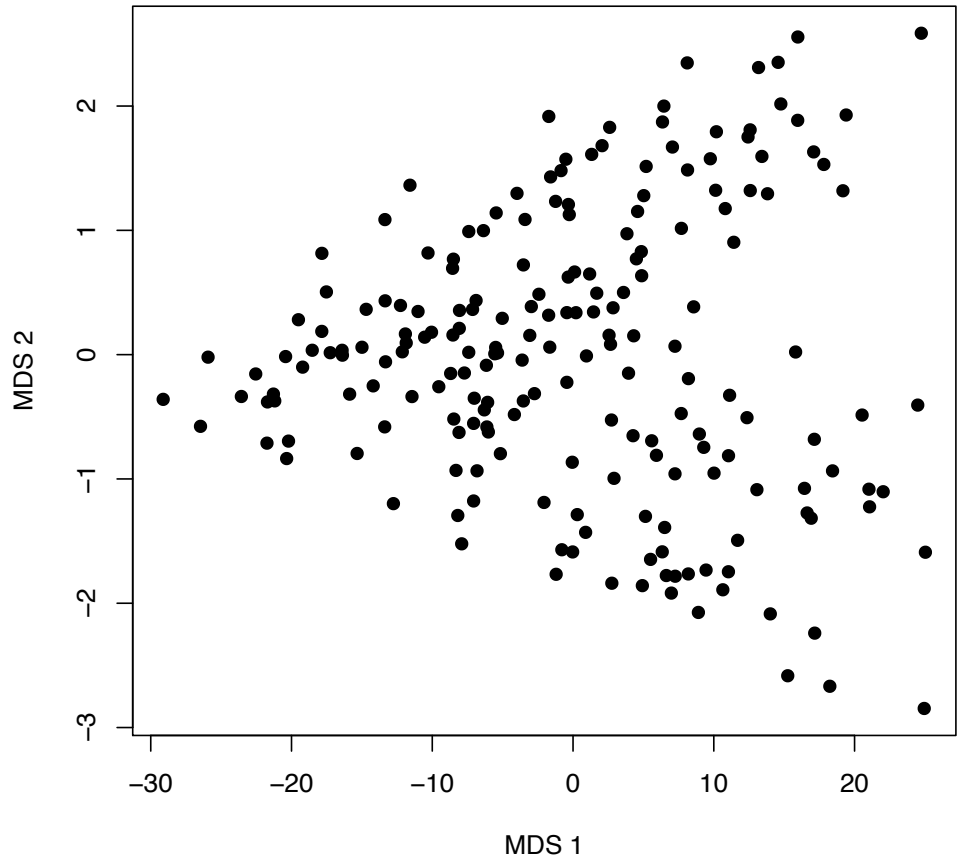
Compare with previous PCA analysis.
MDS solution corresponds to the first 2 PCs as metric scaling was used.



Example: Crabs data

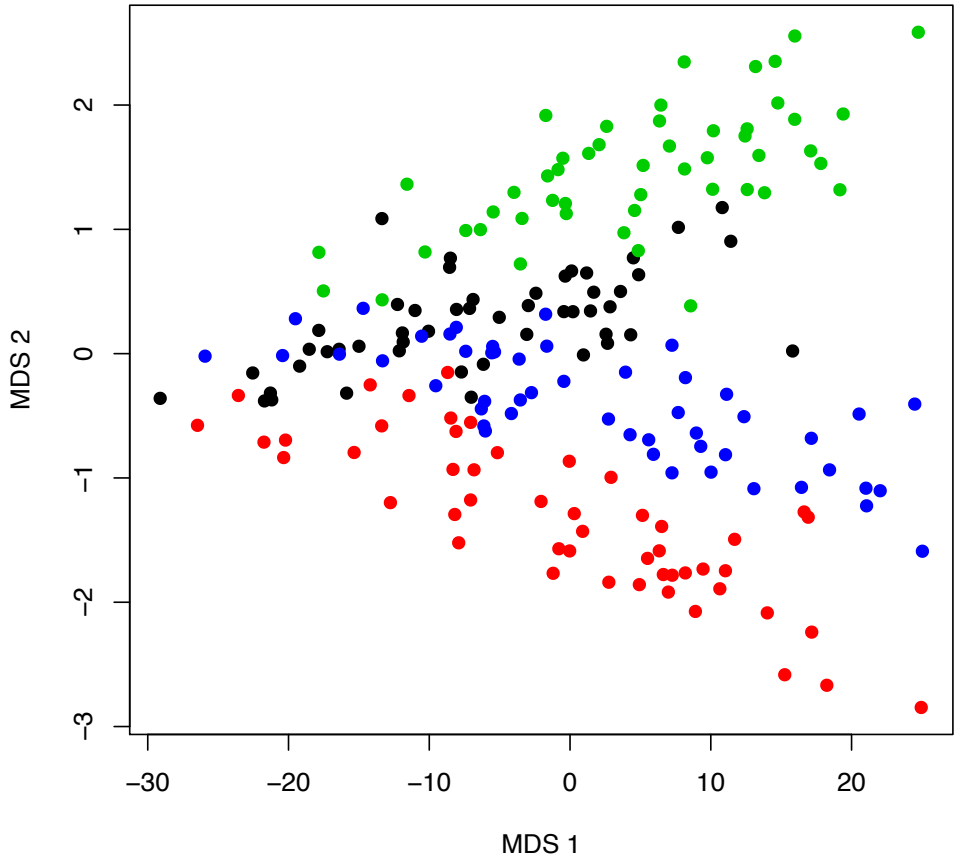
Use Kruskals non-metric multi-dimensional scaling instead.

```
crabsmds <- isoMDS(d= dist(Crabs),k=2)
plot(crabsmds$points, pch=20, cex=2)
```



Example: Crabs data

With grouping information.



Example: Language data

Presence or absence of 2867 homologous traits in 87 Indo-European languages.

```
> X[1:15,1:16]
```

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13
Irish_A	0	0	0	0	1	0	0	0	0	0	0	0	0
Irish_B	0	0	0	0	1	0	0	0	0	0	0	0	0
Welsh_N	0	0	0	1	0	0	0	0	0	0	0	0	0
Welsh_C	0	0	0	1	0	0	0	0	0	0	0	0	0
Breton_List	0	0	0	0	1	0	0	0	0	0	0	0	0
Breton_SE	0	0	0	0	1	0	0	0	0	0	0	0	0
Breton_ST	0	0	0	0	1	0	0	0	0	0	0	0	0
Romanian_List	0	1	0	0	0	0	0	0	0	0	0	0	0
Vlach	0	1	0	0	0	0	0	0	0	0	0	0	0
Italian	0	1	0	0	0	0	0	0	0	0	0	0	0
Ladin	0	1	0	0	0	0	0	0	0	0	0	0	0
Provencal	0	1	0	0	0	0	0	0	0	0	0	0	0
French	0	1	0	0	0	0	0	0	0	0	0	0	0
Walloon	0	1	0	0	0	0	0	0	0	0	0	0	0
French_Creole_C	0	1	0	0	0	0	0	0	0	0	0	0	0

Example: Language data

Using MDS with non-metric scaling.