Distances

In multivariate statistics, distances refer to a single number that summarizes the difference between multivariate observations.
Distances can be defined for:

- Continuous Data
  - genetic distances
  - morphometric distances
- Categorical Data
  - molecular sequences
  - presence/absence data matrices

A distance matrix is always a square $n \times n$ matrix where n is the number of observations in the original dataset

Distance matrices

```
> Y <- cbind(rnorm(5), rnorm(5))
> Y
```

```
           [,1]       [,2]
[1,] -0.9974405  0.7915941
[2,] -1.1910216  0.5255575
[3,]  0.1125328  1.0194546
[4,] -0.4097550 -0.9344067
[5,] -0.6408180 -0.8975576
```

```
> dist(Y)
```

```
          1         2         3         4
2 0.3290123
3 1.1331201 1.3939830
4 1.8233082 1.6558602 2.0224635
5 1.7263873 1.5257722 2.0597265 0.2339828
```

Clustering

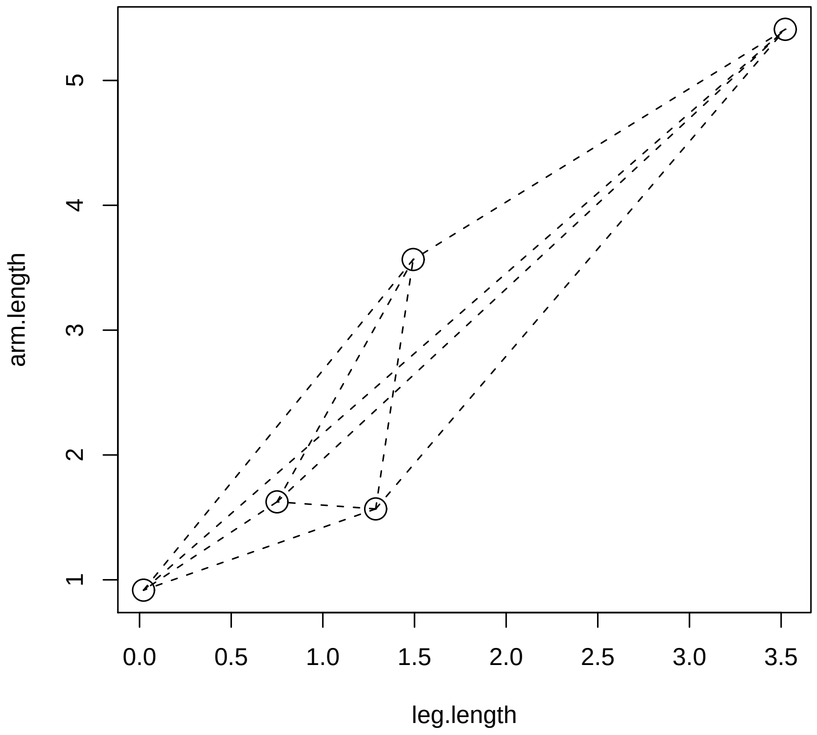
Hierarchical clustering is the means by which distances are most often analysed. There are many slightly different methods. Here we will use WPGMA (Weighted pair-group method using averaging).
Clustering is used to classify

- species
- genetic groups
- types of communities
- soil types

Begin cluster example

```
> library(MASS)
> Z <- data.frame(mvrnorm(5, mu = c(1, 2), Sigma = matrix(c(1,
+     0.7, 0.7, 1), 2, 2)))
> names(Z) <- c("leg.length", "arm.length")
> Z

  leg.length arm.length
1  0.9768720  1.6961592
2  2.3320036  3.4629260
3  1.6416131  2.8530151
4  1.7877215  3.3636897
5 -0.6785844  0.4326824

> plot(Z, cex = 3)
> points(Z, pch = paste(1:5))
> D <- matrix(0, dim(Z)[1], dim(Z)[1])
> for (i in 1:dim(Z)[1]) for (j in i:dim(Z)[1]) {
+     points(Z[c(i, j), ], type = "l", lty = 2)
+     D[i, j] <- sqrt(sum((Z[i, ] - Z[j, ])^2))
+     D[j, i] <- D[i, j]
+ }
```

Cluster example: figure



Distance matrix

```
> D
```

```
          [,1]      [,2]     [,3]     [,4]      [,5]
[1,] 0.0000000 0.1487718 2.197707 1.138336 0.9400500
[2,] 0.1487718 0.0000000 2.049313 1.001752 0.9733489
[3,] 2.1977071 2.0493134 0.000000 1.272697 2.4587166
[4,] 1.1383355 1.0017516 1.272697 0.000000 1.8059823
[5,] 0.9400500 0.9733489 2.458717 1.805982 0.0000000
```

Categorical-based distance

The most common distance measure is based on the proportion of shared states for categorical variables.

```
> x <- c(1, 0, 1, 0, 1, 1, 1, 1, 0)
> y <- c(1, 0, 1, 0, 1, 0, 0, 1, 1)
> x == y
```

```
[1]  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE FALSE
```

```
> Dxy = 1 - sum(as.numeric(x == y))/length(x)
> Dxy
```

```
[1] 0.3333333
```

Performing clustering

On board, but to start

```
> D
```

```
          [,1]      [,2]     [,3]     [,4]      [,5]
[1,] 0.0000000 0.1487718 2.197707 1.138336 0.9400500
[2,] 0.1487718 0.0000000 2.049313 1.001752 0.9733489
[3,] 2.1977071 2.0493134 0.000000 1.272697 2.4587166
[4,] 1.1383355 1.0017516 1.272697 0.000000 1.8059823
[5,] 0.9400500 0.9733489 2.458717 1.805982 0.0000000
```

R implementation: dist

```
> D2 <- dist(Z, upper = T, diag = T)
> D2
```

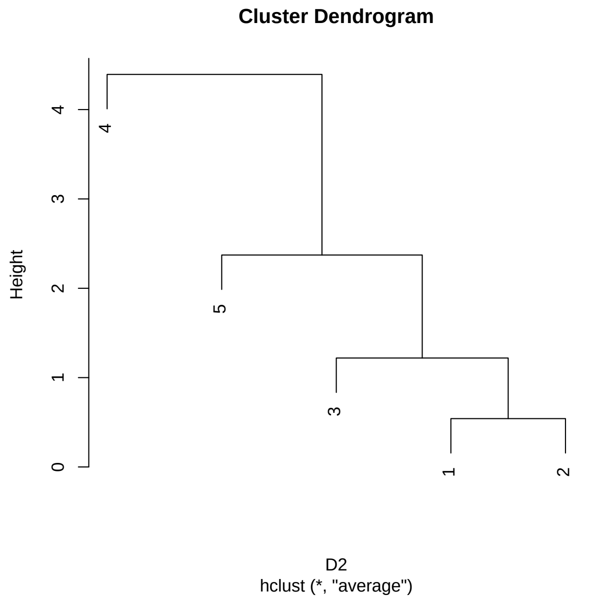
```
          1         2         3         4         5
1 0.0000000 0.1487718 2.1977071 1.1383355 0.9400500
2 0.1487718 0.0000000 2.0493134 1.0017516 0.9733489
3 2.1977071 2.0493134 0.0000000 1.2726966 2.4587166
4 1.1383355 1.0017516 1.2726966 0.0000000 1.8059823
5 0.9400500 0.9733489 2.4587166 1.8059823 0.0000000
```

```
> as.matrix(D2) - D
```

```
  1 2 3 4 5
1 0 0 0 0 0
2 0 0 0 0 0
3 0 0 0 0 0
4 0 0 0 0 0
5 0 0 0 0 0
```

hclust

```
> h.obj <- hclust(D2, method = "average")
> plot(h.obj)
```

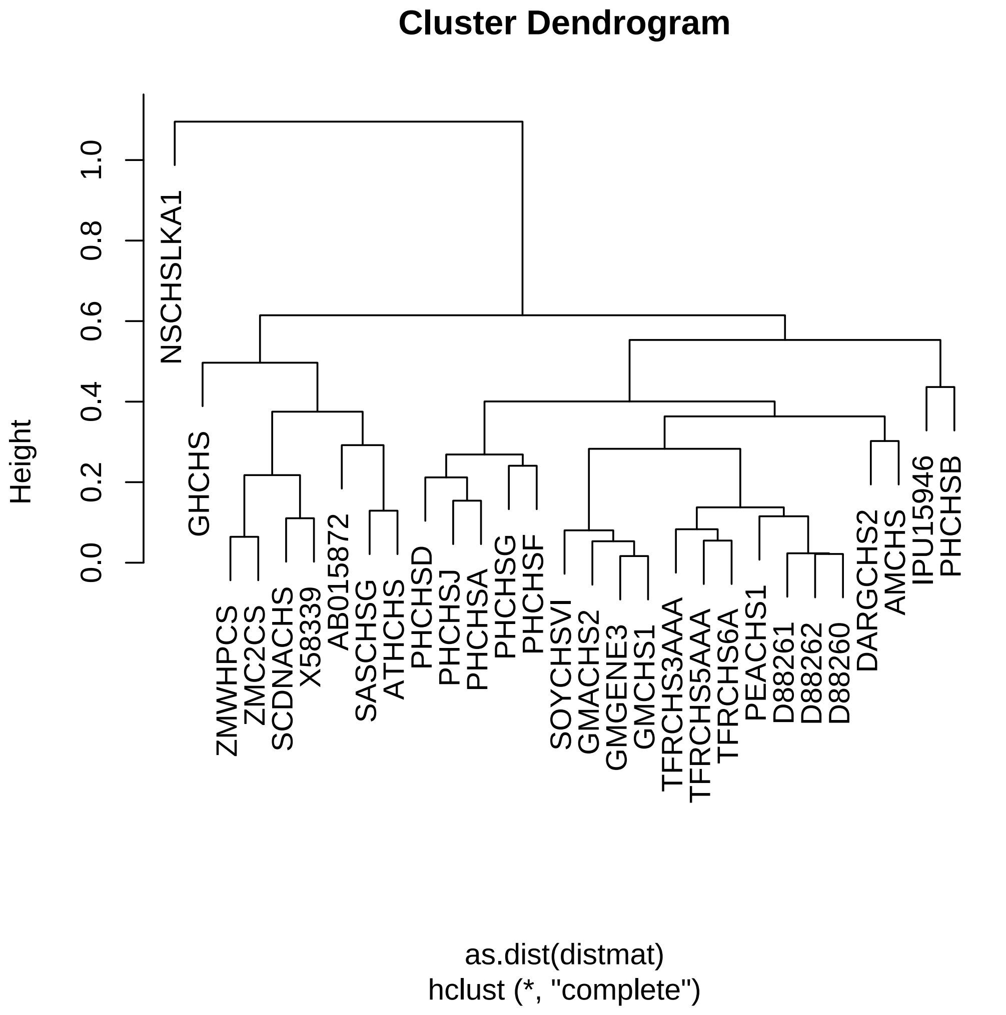


More interesting example

```
> distmat <- read.table("jc.dist", header = F)
> names(distmat) <- c("Gene", paste(distmat[, 1]))
> rownames(distmat) <- distmat[, 1]
> distmat <- as.matrix(distmat[, -1])
> dim(distmat)

[1] 29 29

> distmat[1:10, 1:10]

          D88262 D88261 D88260 TFRCHS5AAA TFRCHS6A TFRCHS3AAA PEACHS1 GMACHS2
D88262    0.0000 0.0234 0.0217     0.1223   0.1283     0.1093  0.1073  0.2656
D88261    0.0234 0.0000 0.0217     0.1243   0.1344     0.1153  0.1123  0.2668
D88260    0.0217 0.0217 0.0000     0.1263   0.1375     0.1163  0.1153  0.2693
TFRCHS5AAA 0.1223 0.1243 0.1263     0.0000   0.0550     0.0831  0.1344  0.2607
TFRCHS6A  0.1283 0.1344 0.1375     0.0550   0.0000     0.0812  0.1314  0.2693
TFRCHS3AAA 0.1093 0.1153 0.1163     0.0831   0.0812     0.0000  0.1143  0.2439
PEACHS1   0.1073 0.1123 0.1153     0.1344   0.1314     0.1143  0.0000  0.2619
GMACHS2   0.2656 0.2668 0.2693     0.2607   0.2693     0.2439  0.2619  0.0000
GMGENE3   0.2619 0.2535 0.2644     0.2656   0.2595     0.2487  0.2607  0.0505
```

```
GMCHS1      0.2619 0.2547 0.2644     0.2619   0.2535     0.2439  0.2595  0.0533
            GMGENE3 GMCHS1
D88262       0.2619 0.2619
D88261       0.2535 0.2547
D88260       0.2644 0.2644
TFRCHS5AAA   0.2656 0.2619
TFRCHS6A     0.2595 0.2535
TFRCHS3AAA   0.2487 0.2439
PEACHS1      0.2607 0.2595
GMACHS2      0.0505 0.0533
GMGENE3      0.0000 0.0165
GMCHS1       0.0165 0.0000
```

cont


Cluster Dendrogram