MULTIVARIATE STATISTICS

Principal Component Analysis and Cluster Analysis
November 6th, 2017

Presented by: Ana Cuciureanu and Shamina S Prova
BIOL 5081: Introduction to Biostatistics
Professor: Dr. Joel Shore

Multivariate Analysis (MVA)

- Complex systems require multiple and different kind of measurements to be taken in order to best describe reality
- MVA is the investigation of many variables, simultaneously, in order to understand the relationships that may exist between variables
- MVA can be as simple as analysing two variables right up to millions

Multivariate Analysis (MVA)

- Is the study of variability and its sources
- Shows the influence of both, wanted and unwanted variability
  - Wanted → the effect of variables on the relationship between data points
  - Unwanted → random variability resulting from experimental features that cannot be controlled
- Used to predict future events

Types of MVA

- Exploratory Data Analysis (EDA)
  - Deeper insight into large, complex data sets
  - i.e. Principle Component Analysis, Cluster Analysis
- Regression Analysis
- Classification
  - Identifies new or existing classes
  - i.e. Cluster Analysis

MVA vs Classical Statistics

- How would you analyze 50 rows and 10 columns of data?
  - Plot columns together two at a time
  - Plot each variable for all samples and look for trends
- This univariate analysis is too simplistic, frustrating and fails to detect the relationship between variants (i.e. covariance and correlation)

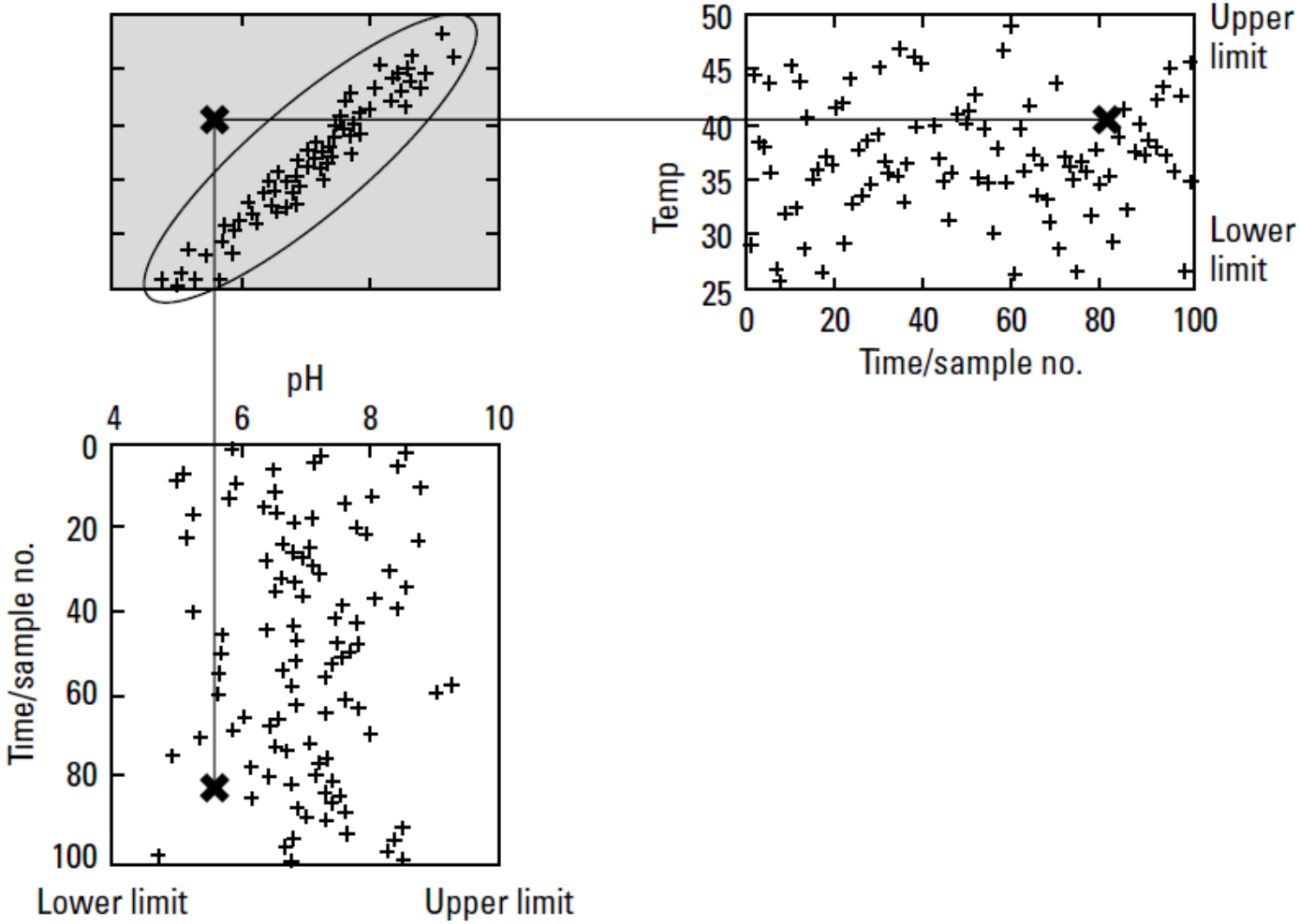
covariance $\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$

correlation $\rho_{XY} = E[(X - \mu_X)(Y - \mu_Y)]/(\sigma_X \sigma_Y)$,

Benefits of MVA

- Identifies variables that contribute most to the overall variability in the data
- Helps isolate those variables that co-vary with each other
- A picture is worth a thousand words and helps understanding the data

Benefits of MVA



Applications of MVA

- Pharmaceutical and biotechnological tests
- Agricultural analysis
- Business intelligence and marketing
- Spectroscopic applications
- Genetics and metabolism
- Etc.

Cell Example

- Imagine you have a dish with a bunch of different cells, but you don’t know how they are characterized
- You decide that the best way to characterize these cell is to measure the mRNA expression of multiple genes
- However, there are too many measurements
- Conclusion: have to use MVA

Cell Example cells = subjects genes = variables

Gene	cell 1	cell 2	cell 3	cell 4	cell 5	cell 6	cell 7	cell 8	cell 9	cell 10
a	12	8	12	8	20	8	8	20	8	24
b	28	28	28	28	0	8	16	12	20	16
c	16	16	16	12	16	16	16	16	16	12
d	20	20	20	20	8	20	20	8	24	8
e	28	24	24	24	4	12	8	20	12	8
f	4	32	12	12	28	0	8	16	16	4
g	18	12	18	12	30	12	12	30	12	36
h	42	42	42	42	0	12	24	18	30	24
i	24	24	24	18	24	24	24	24	24	18
j	30	30	30	30	12	30	30	12	36	12
k	8	12	8	8	20	0	4	28	4	20
l	7	7	7	7	0	2	4	3	5	4
m	8	8	8	6	8	8	8	8	8	6
n	15	15	15	15	12	30	30	12	36	12
o	21	21	21	21	0	6	12	9	15	12

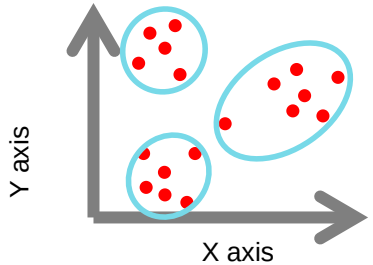
PCA and Cluster Analysis

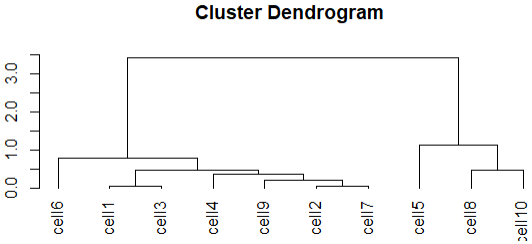
Gene	cell 1	cell 2	cell 3	cell 4	cell 5	cell 6	cell 7	cell 8	cell 9	cell 10
a	12	8	12	8	20	8	8	20	8	24
b	28	28	28	28	0	8	16	12	20	16
c	16	16	16	12	16	16	16	16	16	12
d	20	20	20	20	8	20	20	8	24	8
e	28	24	24	24	4	12	8	20	12	8
f	4	32	12	12	28	0	8	16	16	4
g	18	12	18	12	30	12	12	30	12	36
h	42	42	42	42	0	12	24	18	30	24
i	24	24	24	18	24	24	24	24	24	18
j	30	30	30	30	12	30	30	12	36	12
k	8	12	8	8	20	0	4	28	4	20
l	7	7	7	7	0	2	4	3	5	4
m	8	8	8	6	8	8	8	8	8	6
n	15	15	15	15	12	30	30	12	36	12
o	21	21	21	21	0	6	12	9	15	12

PCA

Cells	Factor 1	Factor 2
cell1	0.93549	0.07226
cell2	0.81307	-0.0181
cell3	0.96315	0.0835
cell4	0.95191	-0.0752
cell5	-0.33566	0.75692
cell6	0.60245	0.0404
cell7	0.8073	0.0129
cell8	-0.01531	0.95305
cell9	0.8369	-0.07732
cell10	0.18234	0.85819

Cluster Analysis





PRINCIPAL COMPONENT ANALYSIS

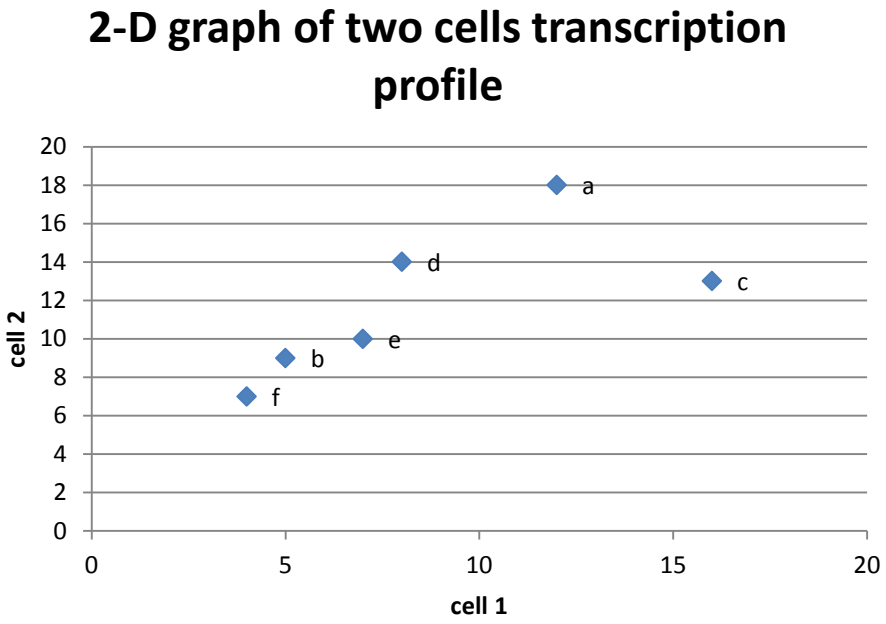
Reduction of dimension

1 dimension number line

Transcription from single cell	
Gene	mRNA count
a	12
b	5
c	16
d	8
e	7
f	4

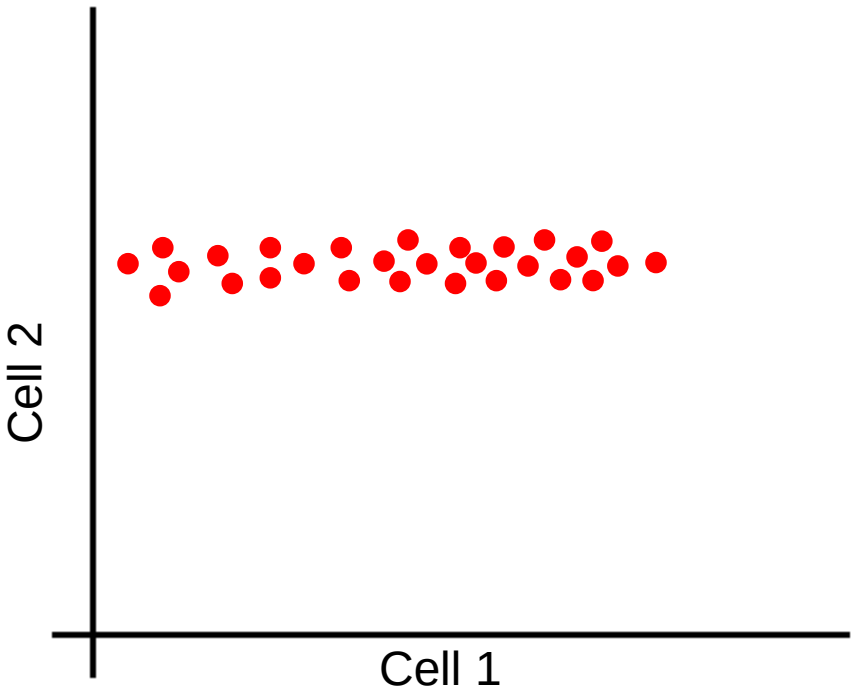
Two dimension graph

Gene	Cell1	Cell2
a	12	18
b	5	9
c	16	13
d	8	14
e	7	10
f	4	7

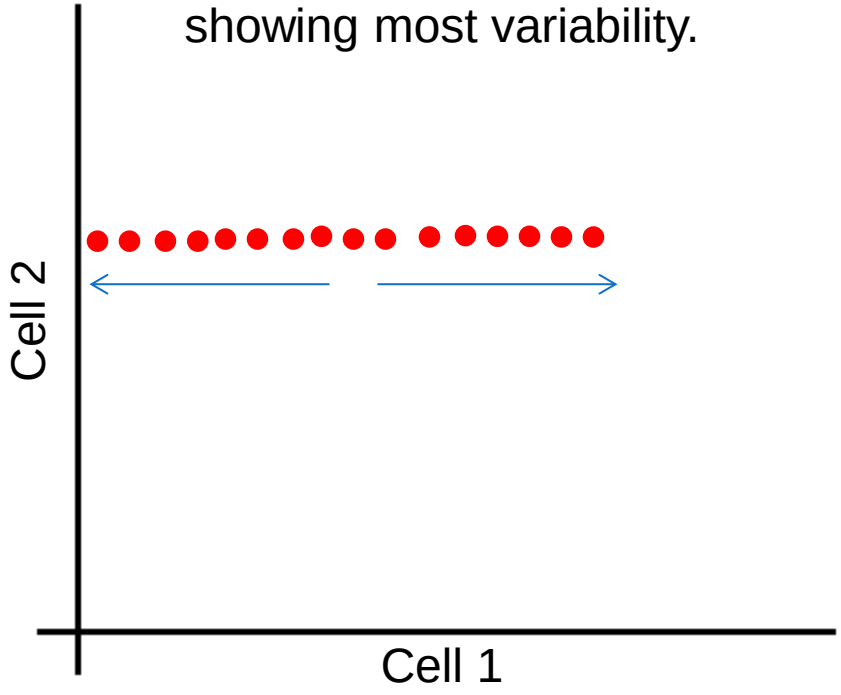


- So for 3 cells….it requires a 3-D data graph.
- For 4 cells…4 dimensional…which is not possible to draw on paper
- For 1000 cells…..1000-D (Impossible!)

Principal Component Determination

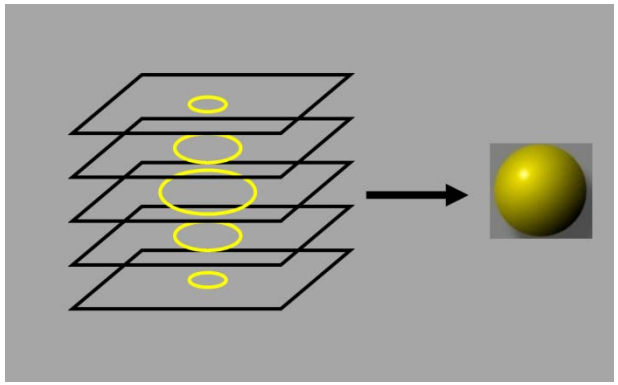


Flattening of data. One axis showing most variability.



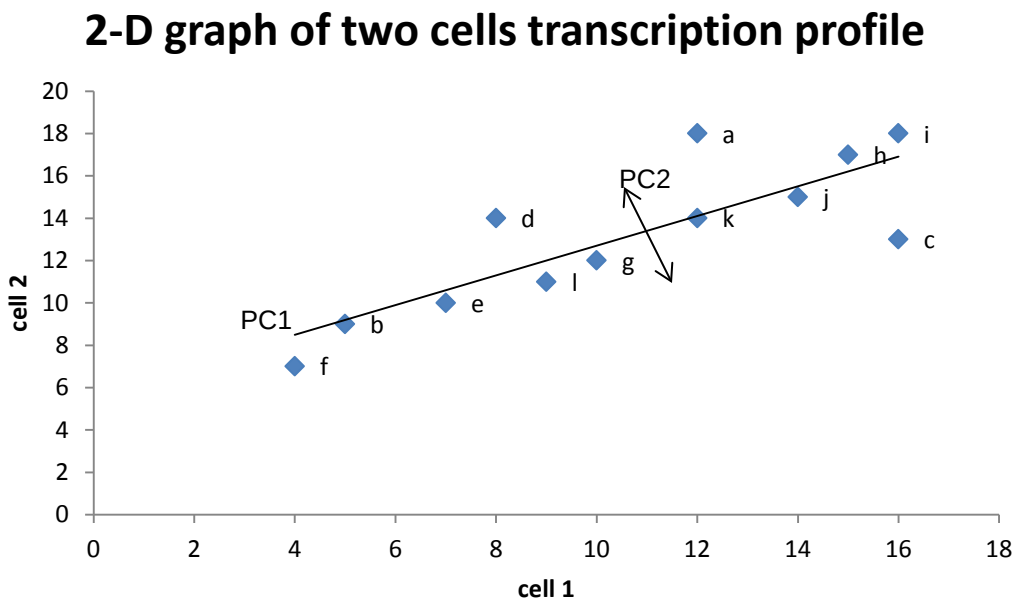
Principal Component Analysis

- PCA compresses(flattens) multidimensional data(multiple cell) into 2 or 3 dimensions which provides meaningful interpretation about the maximum variance in the data set.
- Flattening a Z stack of microscope images to make a 2-D image for paper.



Principal Component Analysis

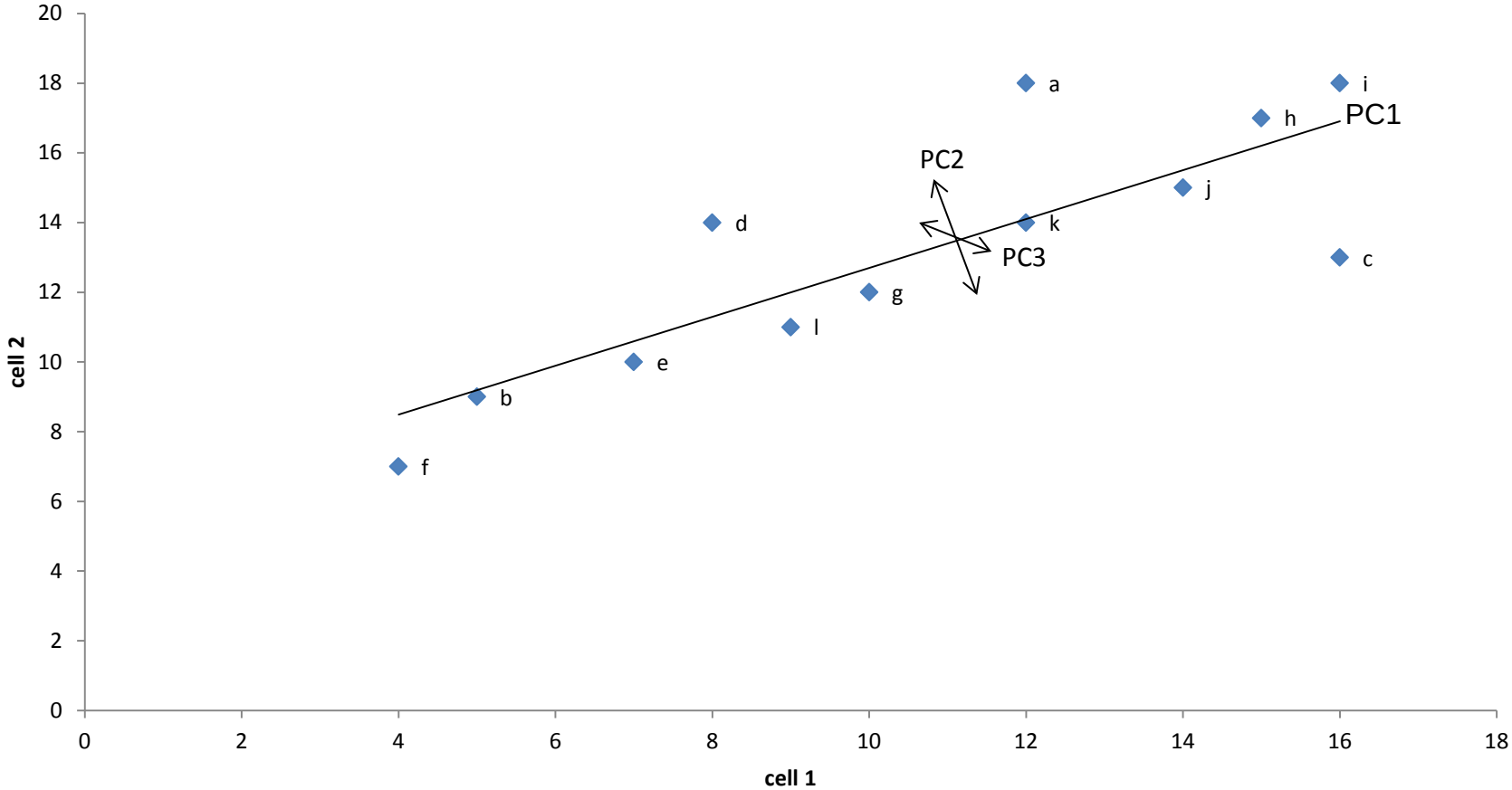
Gene	Cell1	Cell2
a	12	18
b	5	9
c	16	13
d	8	14
e	7	10
f	4	7
g	10	12
h	15	17
i	16	18
j	14	15
k	12	14
l	9	11



PC1: Most variation axis

PC2: the 2nd most variation axis

Principal Component Analysis

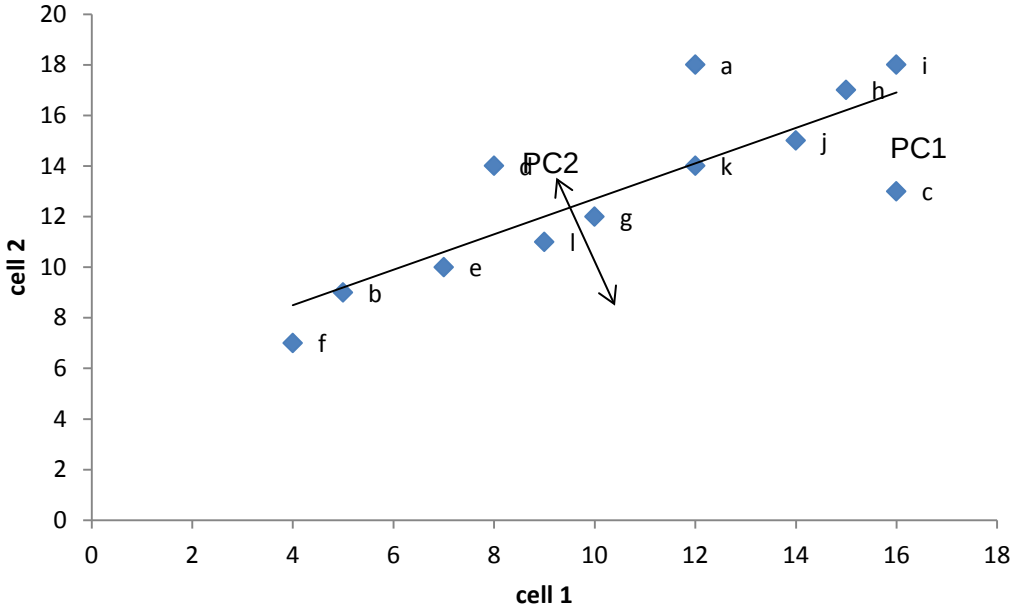


For 500 cells….500 Principal component!

Variability extent on each PC

Loading: Influence of each gene on the PC.

Eigenvalue and Eigenvector: an array of loading for a PC with direction of the influence.



Gene	Cell1	Cell2
a	12	18
b	5	9
c	16	13
d	8	14
e	7	10
f	4	7
g	10	12

Gene	Influence on PC1	In value PC1	Influence on PC2	In value PC2
a	Medium	5	High	3
b	High	-9	Low	-0.1
c	High	9	High	-3.5
d	Low	-3	High	2
e	Medium	-6	Low	-0.5
f	High	-11	Medium	-1
g	Low	-0.5	Low	-0.5

Variability Scoring

Gene	Cell1	Cell2
a	12	18
b	5	9
c	16	13
d	8	14
e	7	10
f	4	7
g	10	12

Gene	Influence on PC1	In value PC1	Influence on PC2	In value PC2
a	Medium	5	High	3
b	High	-9	Low	-0.1
c	High	9	High	-3.5
d	Low	-3	High	2
e	Medium	-6	Low	-0.5
f	High	-11	Medium	-1
g	Low	-0.5	Low	-0.5

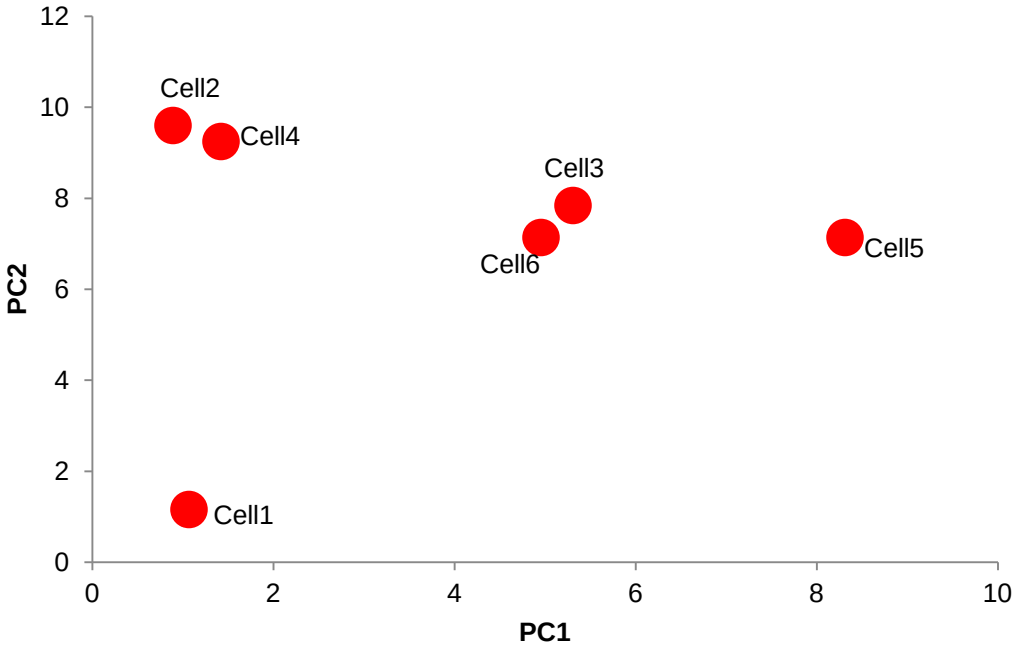
Score cell based on transcription level and influence on each principal component:

Cell1 PC1 =Σ(no. of expression of gene * respective influence on PC1)
=(12*5)+(5*-9)+(16*9)+(8*-3)+…….
=1

Cell1 PC2 =(18*3)+(9*-0.1)+(13*-3.5)+(14*2)+……
=1

Plotting PC2 against PC1

Cell	PC1	PC2
Cell 1	1	1
Cell 2	0.8	9.5
Cell 3	5.5	7.5
Cell 4	1.5	9



Mathematical representation

- $X = [\quad]_{nxm}$ where X is the data matrix, with n no. of samples and m no. of measurements.
- PCA=Eigendecomposition, $X^T X = W$ where W is the eigenvalues with eigenvectors (mXm matrix) and X^T is the X transpose matrix.
- T=XW where T is the score (nXm matrix).

Characteristics of W is such that each column is a PC and the eigenvalues are arranged in descending order.

Assumptions

1. Linearity
2. Correlation among the variables
3. Large variance have more important dynamics
4. Sample size: 150+ cases.
5. All outliers should be removed
6. Components are uncorrelated

PCA BY R

Data

Transcription level of 15 genes in 10 different cells.

Gene	cell 1	cell 2	cell 3	cell 4	cell 5	cell 6	cell 7	cell 8	cell 9	cell 10
a	12	8	12	8	20	8	8	20	8	24
b	28	28	28	28	0	8	16	12	20	16
c	16	16	16	12	16	16	16	16	16	12
d	20	20	20	20	8	20	20	8	24	8
e	28	24	24	24	4	12	8	20	12	8
f	4	32	12	12	28	0	8	16	16	4
g	18	12	18	12	30	12	12	30	12	36
h	42	42	42	42	0	12	24	18	30	24
i	24	24	24	18	24	24	24	24	24	18
j	30	30	30	30	12	30	30	12	36	12
k	8	12	8	8	20	0	4	28	4	20
l	7	7	7	7	0	2	4	3	5	4
m	8	8	8	6	8	8	8	8	8	6
n	15	15	15	15	12	30	30	12	36	12
o	21	21	21	21	0	6	12	9	15	12

Data in R

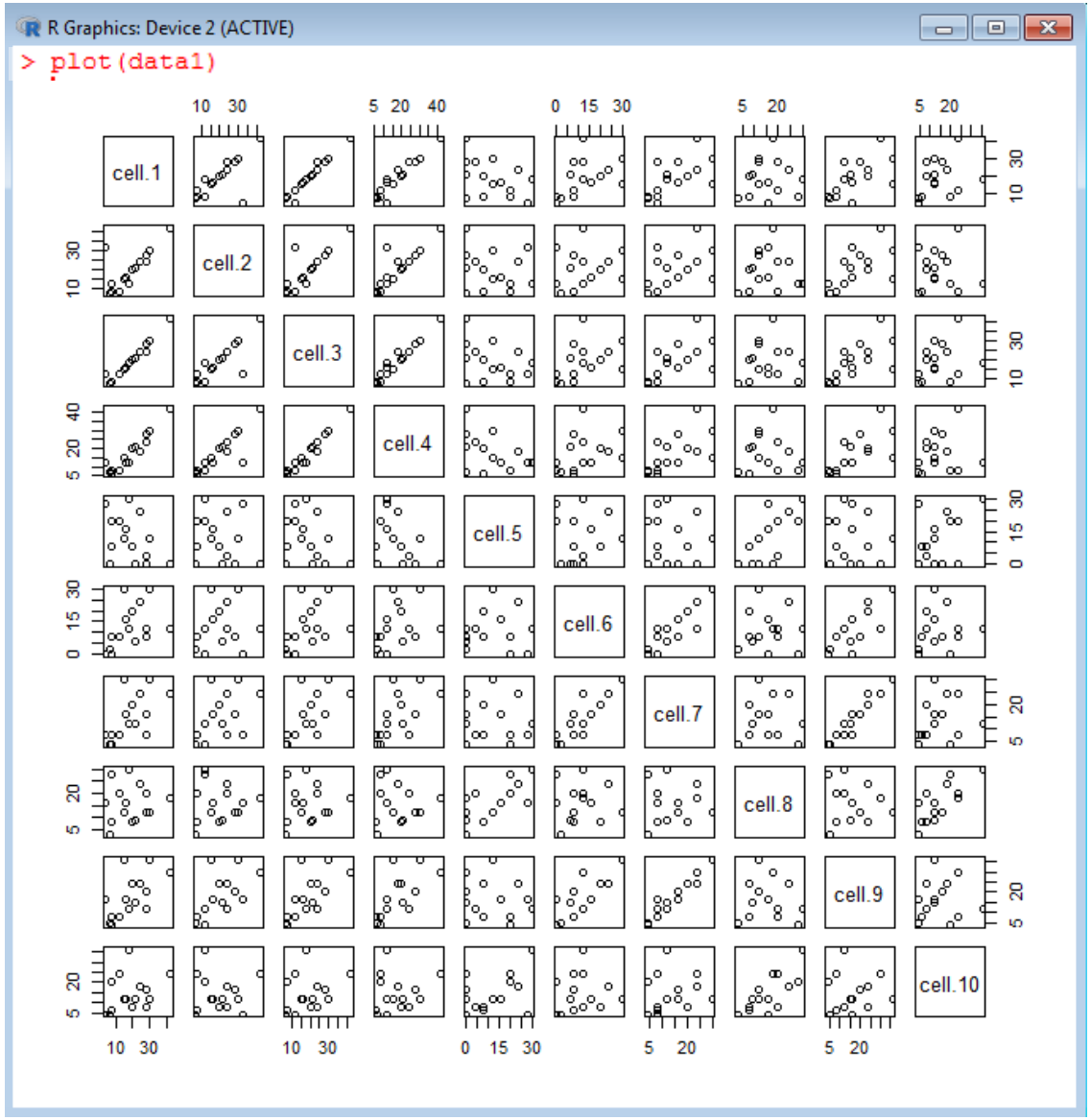
RGui (32-bit)

File Edit View Misc Packages Windows Help

R Console

```
> data1 <-read.csv(file.choose(),header=TRUE)
> data1
   cell.1 cell.2 cell.3 cell.4 cell.5 cell.6 cell.7 cell.8 cell.9 cell.10
1      12      8     12      8     20      8      8     20      8     24
2      28     28     28     28      0      8     16     12     20     16
3      16     16     16     12     16     16     16     16     16     12
4      20     20     20     20      8     20     20      8     24      8
5      28     24     24     24      4     12      8     20     12      8
6       4     32     12     12     28      0      8     16     16      4
7      18     12     18     12     30     12     12     30     12     36
8      42     42     42     42      0     12     24     18     30     24
9      24     24     24     18     24     24     24     24     24     18
10     30     30     30     30     12     30     30     12     36     12
11      8     12      8      8     20      0      4     28      4     20
12      7      7      7      7      0      2      4      3      5      4
13      8      8      8      6      8      8      8      8      8      6
14     15     15     15     15     12     30     30     12     36     12
15     21     21     21     21      0      6     12      9     15     12
>
```

Correlation among cells

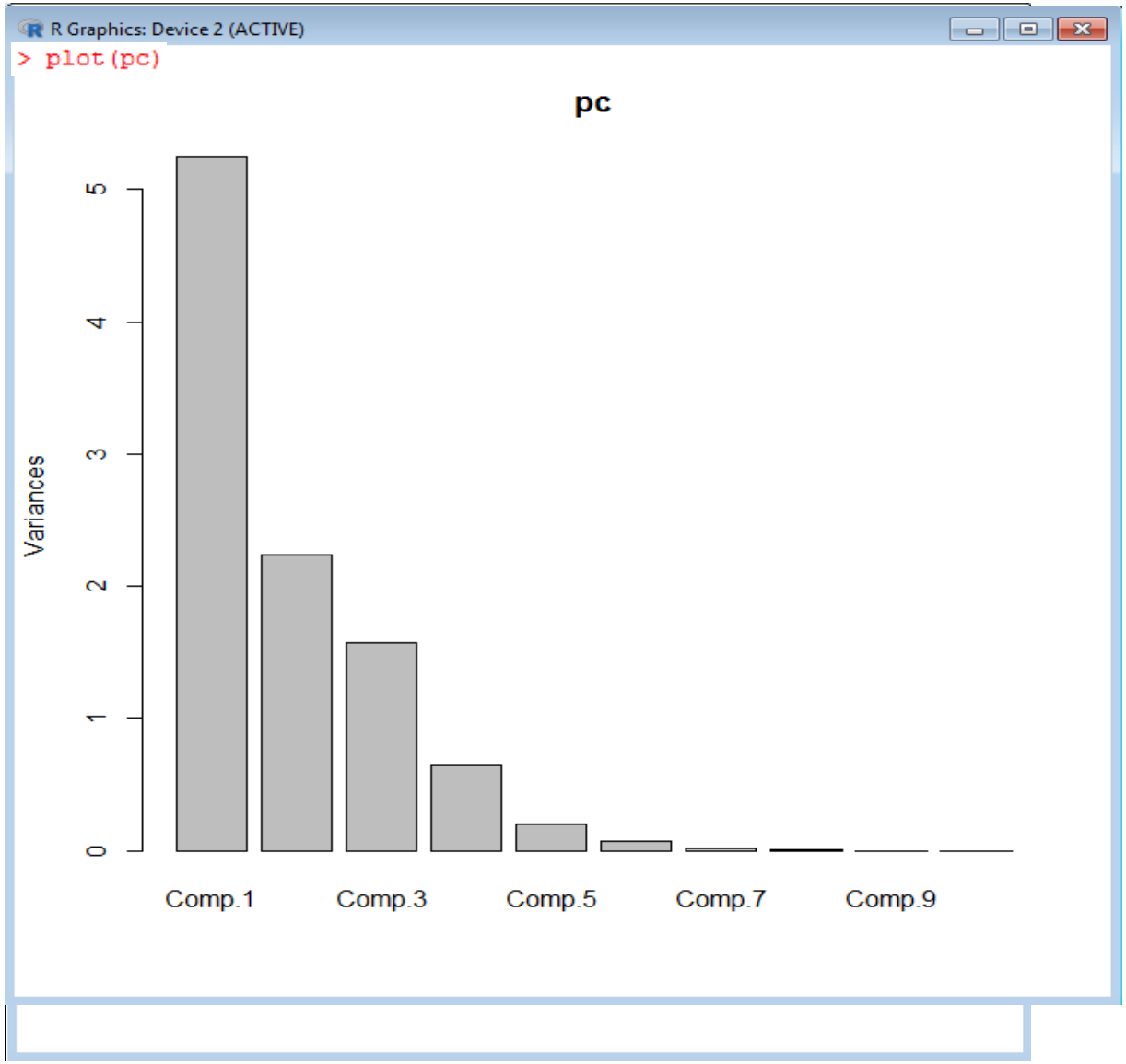


PCA summary

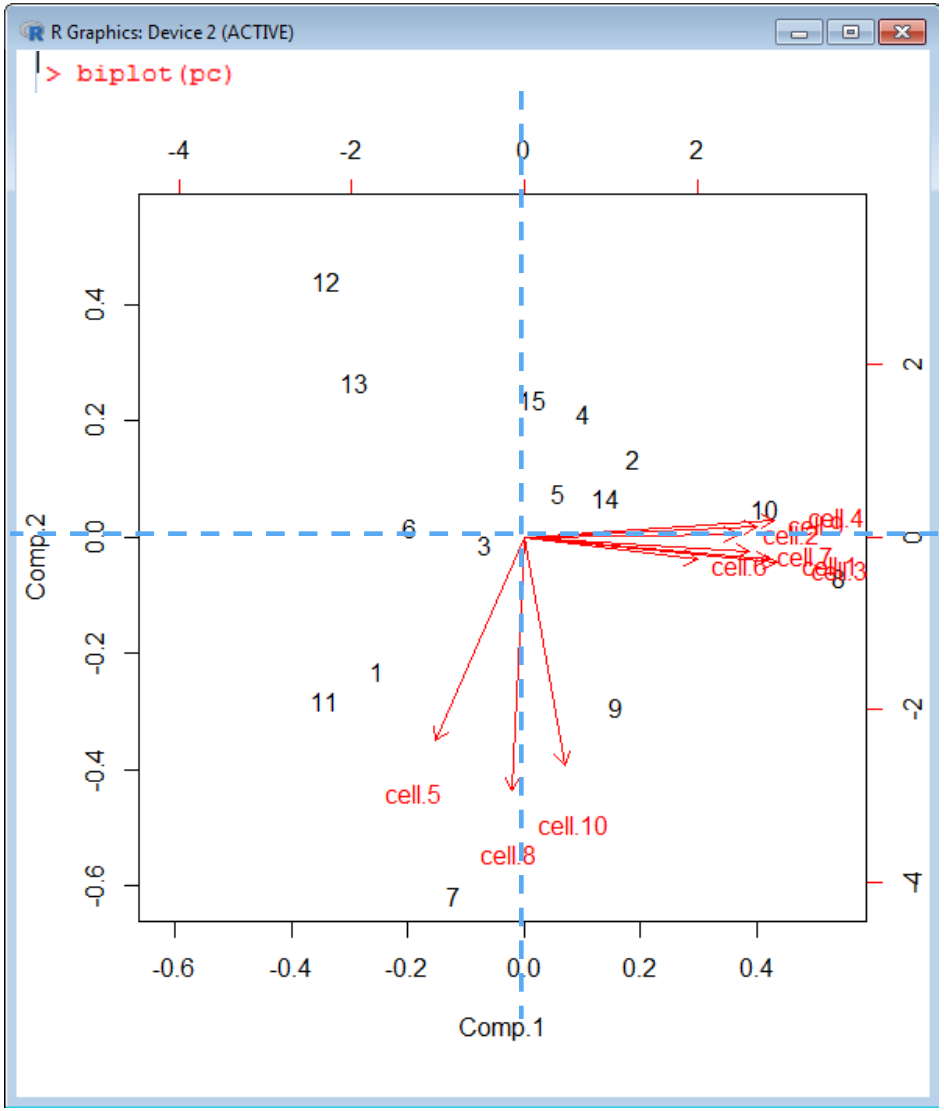
R Console

```
> pc <-princomp(data1, score=TRUE, cor=TRUE)
> summary(pc)
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
Standard deviation     2.2911491 1.4975273 1.2533352 0.80823993 0.44571170
Proportion of Variance 0.5249364 0.2242588 0.1570849 0.06532518 0.01986589
Cumulative Proportion  0.5249364 0.7491952 0.9062801 0.97160530 0.99147119
                           Comp.6      Comp.7       Comp.8       Comp.9
Standard deviation     0.26074355 0.121463216 0.0504736140 1.233474e-08
Proportion of Variance 0.00679872 0.001475331 0.0002547586 1.521458e-17
Cumulative Proportion  0.99826991 0.999745241 1.0000000000 1.000000e+00
                            Comp.10
Standard deviation     3.756283e-09
Proportion of Variance 1.410966e-18
Cumulative Proportion  1.000000e+00
>
```

Scree plot


R Graphics: Device 2 (ACTIVE)
> plot(pc)
pc
Variances
0
1
2
3
4
5
Comp.1
Comp.3
Comp.5
Comp.7
Comp.9

Graphical representation PC2 vs PC1



Loadings by different components on each cell

R Console

```
> pc$loading

Loadings:
        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10
cell.1  -0.399        -0.244  0.243 -0.324 -0.251 -0.108 -0.135  0.722
cell.2  -0.348        -0.251 -0.635  0.110        -0.142 -0.203         0.578
cell.3  -0.411        -0.243               -0.359 -0.206 -0.228 -0.539 -0.502
cell.4  -0.408        -0.272                       0.522  0.687
cell.5   0.145 -0.509  0.296 -0.481        -0.526  0.114  0.182  0.198 -0.189
cell.6  -0.282         0.565  0.224 -0.444 -0.196  0.235        -0.292  0.407
cell.7  -0.367         0.413         0.212  0.187 -0.651  0.427
cell.8         -0.632 -0.131        -0.480  0.563                       -0.108
cell.9  -0.380         0.354 -0.140  0.307  0.363  0.385 -0.430  0.200 -0.332
cell.10        -0.571 -0.156  0.472  0.557                              0.295

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
SS loadings       1.0    1.0    1.0    1.0    1.0    1.0    1.0    1.0    1.0
Proportion Var    0.1    0.1    0.1    0.1    0.1    0.1    0.1    0.1    0.1
Cumulative Var    0.1    0.2    0.3    0.4    0.5    0.6    0.7    0.8    0.9
               Comp.10
SS loadings        1.0
Proportion Var     0.1
Cumulative Var     1.0
>
```

Scores of genes

R Console

```
> pc$score
          Comp.1      Comp.2      Comp.3       Comp.4      Comp.5       Comp.6
 [1,]  2.2450281 -1.33064649 -0.05607516  0.717942907  0.25558131 -0.216892660
 [2,] -1.6763363  0.78306085 -1.41236564  0.263598637  0.33668259  0.048727455
 [3,]  0.6001011 -0.07550643  0.74783680 -0.004553757 -0.28540144 -0.138325688
 [4,] -0.9073793  1.23961560  0.91752584  0.083465319 -0.02369290 -0.194324944
 [5,] -0.5249865  0.43814682 -1.40227847 -0.053354720 -1.33574610  0.074620735
 [6,]  1.7393582  0.08976140 -0.11102158 -2.774060033  0.40161486 -0.065391029
 [7,]  1.0710035 -3.57276903 -0.10083988  0.721943002  0.31734920 -0.261368724
 [8,] -4.8110431 -0.40220802 -2.13527560  0.040426596  0.43900112  0.137061450
 [9,] -1.3963871 -1.69005894  1.10502806 -0.361801994 -0.49412493 -0.143518265
[10,] -3.6576076  0.28262411  1.35956162 -0.229773380 -0.10156213 -0.227517149
[11,]  3.0526422 -1.63480840 -0.83622632 -0.131919745 -0.17445924  0.611807994
[12,]  3.0257239  2.56096415 -0.32800069  0.598356697  0.18320480 -0.083773537
[13,]  2.5965892  1.53904607  0.39064554  0.352694480 -0.07667795 -0.133133111
[14,] -1.2477236  0.39708304  2.91239613  0.401851333  0.27270747  0.587467015
[15,] -0.1089829  1.37569528 -1.05091065  0.375184657  0.28552333  0.004560458
           Comp.7       Comp.8        Comp.9       Comp.10
 [1,]  0.08149097 -0.020557644 -6.611804e-16  8.329112e-15
 [2,] -0.04781310 -0.007876108  9.851823e-16  5.418504e-16
 [3,] -0.18714218 -0.004684563  2.136170e-15 -4.138718e-15
 [4,]  0.09902090  0.063162225 -1.851207e-15 -1.380708e-14
 [5,]  0.16893356 -0.064006881 -3.003064e-15  2.235843e-14
 [6,]  0.06179806 -0.035633258 -1.678487e-15  1.049486e-14
 [7,]  0.12261938 -0.031471151 -1.088675e-15  1.219017e-14
 [8,] -0.07133674 -0.012448848  1.295588e-15  7.240844e-16
 [9,] -0.28033036 -0.007661531  3.049995e-15 -6.357633e-15
[10,]  0.14891427  0.094108652 -2.849426e-15 -2.077819e-14
[11,] -0.01517407  0.111578301  1.176871e-15 -2.976349e-14
[12,] -0.01252765 -0.001016998  4.031809e-16  2.808084e-16
[13,] -0.09395401 -0.001707596  1.041708e-15 -1.847887e-15
[14,]  0.06155226 -0.076194862 -1.406136e-15  2.163233e-14
[15,] -0.03605128 -0.005589738  7.449214e-16  4.548296e-16
>
```

PCR IN SAS

Data with code in SAS

Editor - Untitled1 *

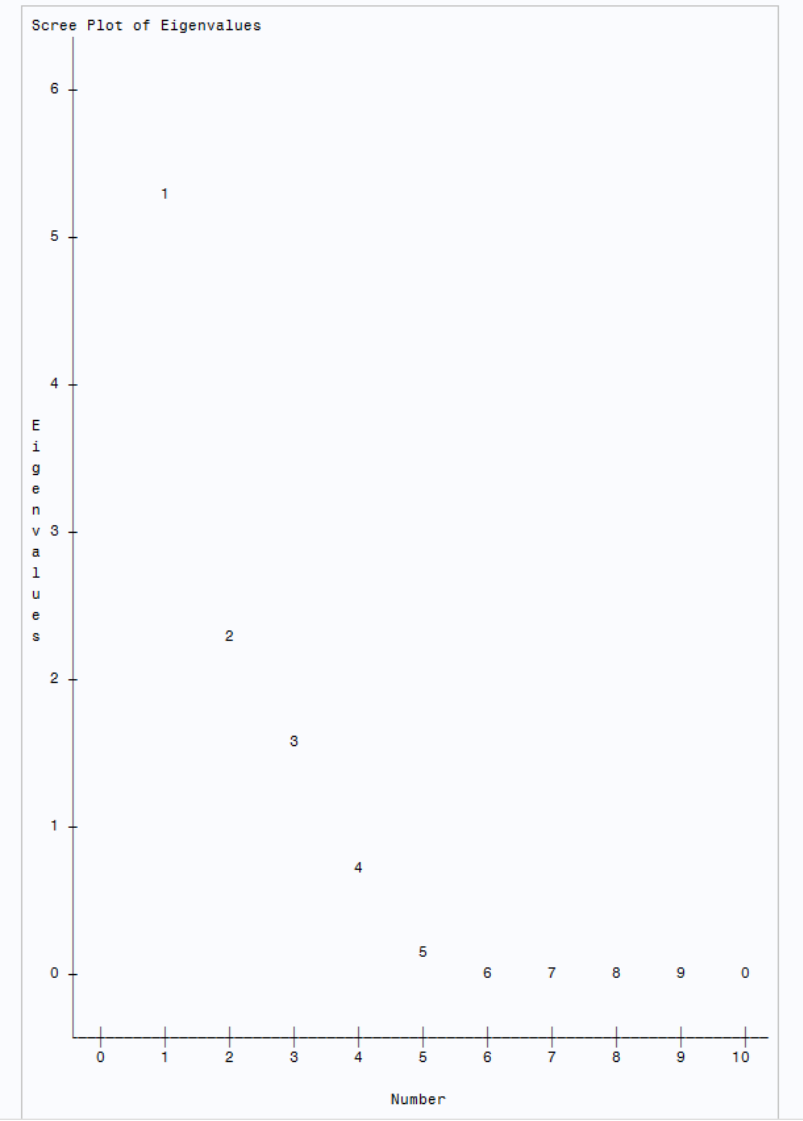
```
option linesize=80;
data pc;
title 'GENE TRANSCRIPTION';
   input genes$ cell1 cell2 cell3 cell4 cell5 cell6 cell7 cell8 cell9 cell10;
   cards;
a   12  8   12  8   20  8   8   20  8   24
b   28  28  28  28  0   8   16  12  20  16
c   16  16  16  12  16  16  16  16  16  12
d   20  20  20  20  8   20  20  8   24  8
e   28  24  24  24  4   12  8   20  12  8
f   4   32  12  12  28  0   8   16  16  4
g   18  12  18  12  30  12  12  30  12  36
h   42  42  42  42  0   12  24  18  30  24
i   24  24  24  18  24  24  24  24  24  18
j   30  30  30  30  12  30  30  12  36  12
k   8   12  8   8   20  0   4   28  4   20
l   7   7   7   7   0   2   4   3   5   4
m   8   8   8   6   8   8   8   8   8   6
n   15  15  15  15  12  30  30  12  36  12
o   21  21  21  21  0   6   12  9   15  12

;
proc corr;
proc factor data=pc method=prin scree;
var cell1 cell2 cell3 cell4 cell5 cell6 cell7 cell8 cell9 cell10;
run;
```

Correlation among cells

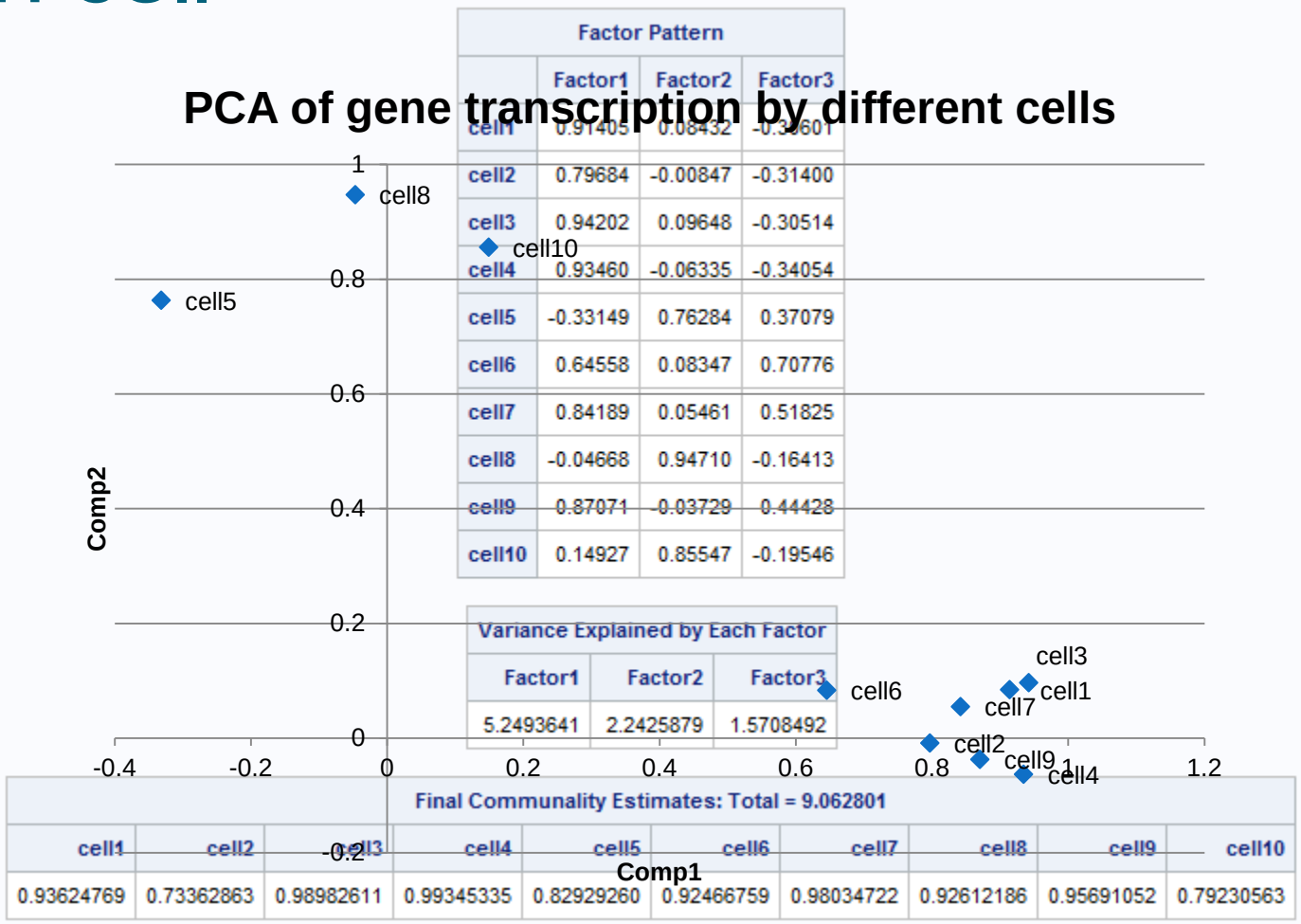
Pearson Correlation Coefficients, N = 15
Prob > |r| under H0: Rho=0

	cell1	cell2	cell3	cell4	cell5	cell6	cell7	cell8	cell9	cell10
cell1	1.00000	0.71600 0.0027	0.97531 <.0001	0.94440 <.0001	-0.42357 0.1157	0.44756 0.0944	0.60767 0.0163	0.09305 0.7415	0.60825 0.0161	0.30825 0.2637
cell2	0.71600 0.0027	1.00000	0.83572 0.0001	0.86697 <.0001	-0.18724 0.5040	0.18830 0.5015	0.49264 0.0621	0.03796 0.8931	0.61913 0.0139	-0.01082 0.9695
cell3	0.97531 <.0001	0.83572 0.0001	1.00000	0.97389 <.0001	-0.34687 0.2053	0.41133 0.1277	0.63632 0.0108	0.08662 0.7589	0.66668 0.0066	0.28612 0.3012
cell4	0.94440 <.0001	0.86697 <.0001	0.97389 <.0001	1.00000	-0.47358 0.0746	0.35165 0.1987	0.60247 0.0175	-0.04359 0.8774	0.67265 0.0060	0.14077 0.6168
cell5	-0.42357 0.1157	-0.18724 0.5040	-0.34687 0.2053	-0.47358 0.0746	1.00000	0.04382 0.8768	-0.06613 0.8149	0.68216 0.0051	-0.11728 0.6772	0.39158 0.1489
cell6	0.44756 0.0944	0.18830 0.5015	0.41133 0.1277	0.35165 0.1987	0.04382 0.8768	1.00000	0.89875 <.0001	-0.04721 0.8673	0.82253 0.0002	0.05032 0.8586
cell7	0.60767 0.0163	0.49264 0.0621	0.63632 0.0108	0.60247 0.0175	-0.06613 0.8149	0.89875 <.0001	1.00000	-0.08824 0.7545	0.96998 <.0001	0.10840 0.7006
cell8	0.09305 0.7415	0.03796 0.8931	0.08662 0.7589	-0.04359 0.8774	0.68216 0.0051	-0.04721 0.8673	-0.08824 0.7545	1.00000	-0.15554 0.5799	0.74906 0.0013
cell9	0.60825 0.0161	0.61913 0.0139	0.66668 0.0066	0.67265 0.0060	-0.11728 0.6772	0.82253 0.0002	0.96998 <.0001	-0.15554 0.5799	1.00000	0.00124 0.9965
cell10	0.30825 0.2637	-0.01082 0.9695	0.28612 0.3012	0.14077 0.6168	0.39158 0.1489	0.05032 0.8586	0.10840 0.7006	0.74906 0.0013	0.00124 0.9965	1.00000

PCA summary


Scree Plot of Eigenvalues
Eigenvalues
Number
0
1
2
3
4
5
6
7
8
9
10

Loadings by different components on each cell



Factor Pattern

	Factor1	Factor2	Factor3
cell1	0.91405	0.08432	-0.39601
cell2	0.79684	-0.00847	-0.31400
cell3	0.94202	0.09648	-0.30514
cell4	0.93460	-0.06335	-0.34054
cell5	-0.33149	0.76284	0.37079
cell6	0.64558	0.08347	0.70776
cell7	0.84189	0.05461	0.51825
cell8	-0.04668	0.94710	-0.16413
cell9	0.87071	-0.03729	0.44428
cell10	0.14927	0.85547	-0.19546

Variance Explained by Each Factor

Factor1	Factor2	Factor3
5.2493641	2.2425879	1.5708492

Final Communality Estimates: Total = 9.062801

cell1	cell2	cell3	cell4	cell5	cell6	cell7	cell8	cell9	cell10
0.93624769	0.73362863	0.98982611	0.99345335	0.82929260	0.92466759	0.98034722	0.92612186	0.95691052	0.79230563

Orthogonal rotation

```
proc factor data=pc method=prin scree n=3 out=scores rotate=varimax;
var cell1 cell2 cell3 cell4 cell5 cell6 cell7 cell8 cell9 cell10;
run;
```

GENE TRANSCRIPTION

The FACTOR Procedure
Rotation Method: Varimax

Orthogonal Transformation Matrix

	1	2	3
1	0.81452	0.57959	-0.02521
2	-0.01391	0.06296	0.99792
3	-0.57997	0.81247	-0.05934

Rotated Factor Pattern

	Factor1	Factor2	Factor3
cell1	0.92082	0.28646	0.07926
cell2	0.83127	0.20619	-0.00991
cell3	0.94292	0.30414	0.09064
cell4	0.95963	0.26102	-0.06657
cell5	-0.49567	0.15715	0.74760
cell6	0.11419	0.95446	0.02502
cell7	0.38440	0.91246	0.00252
cell8	0.04399	-0.10078	0.95605
cell9	0.45205	0.86327	-0.08553
cell10	0.22304	-0.01843	0.86152

Variance Explained by Each Factor

Factor1	Factor2	Factor3
4.0114431	2.8092250	2.2421331

Final Communality Estimates: Total = 9.062801

cell1	cell2	cell3	cell4	cell5	cell6	cell7	cell8	cell9	cell10
0.93624769	0.73362863	0.98982611	0.99345335	0.82929260	0.92466759	0.98034722	0.92612186	0.95691052	0.79230563

Limitation of PCA

- Requires numeric data for analysis
- 150+ data needed to get a representative factor trend.
- Loss of information due to dimension reduction
- Analysis is non conclusive. Needs explanatory factor analysis or cluster analysis to explain overall trend.

CLUSTER ANALYSIS

Cluster Analysis

- An unsupervised learning tool
- It breaks down a large data set into smaller groups (i.e. clusters) where observations within a group are more similar than observations from other groups.

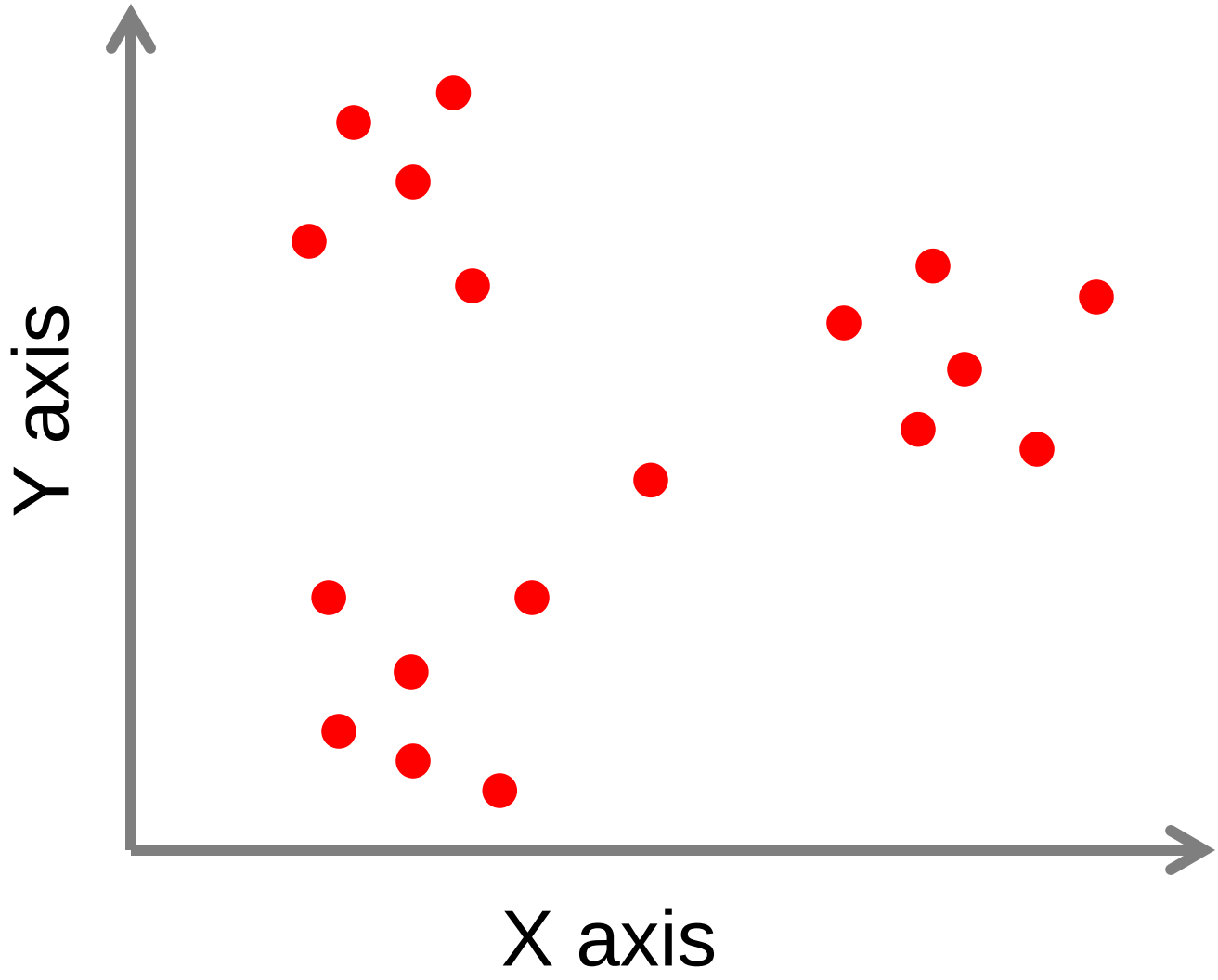
Algorithms

- Hierarchical Cluster
- Non-Hierarchical Cluster (aka K-Means Cluster)

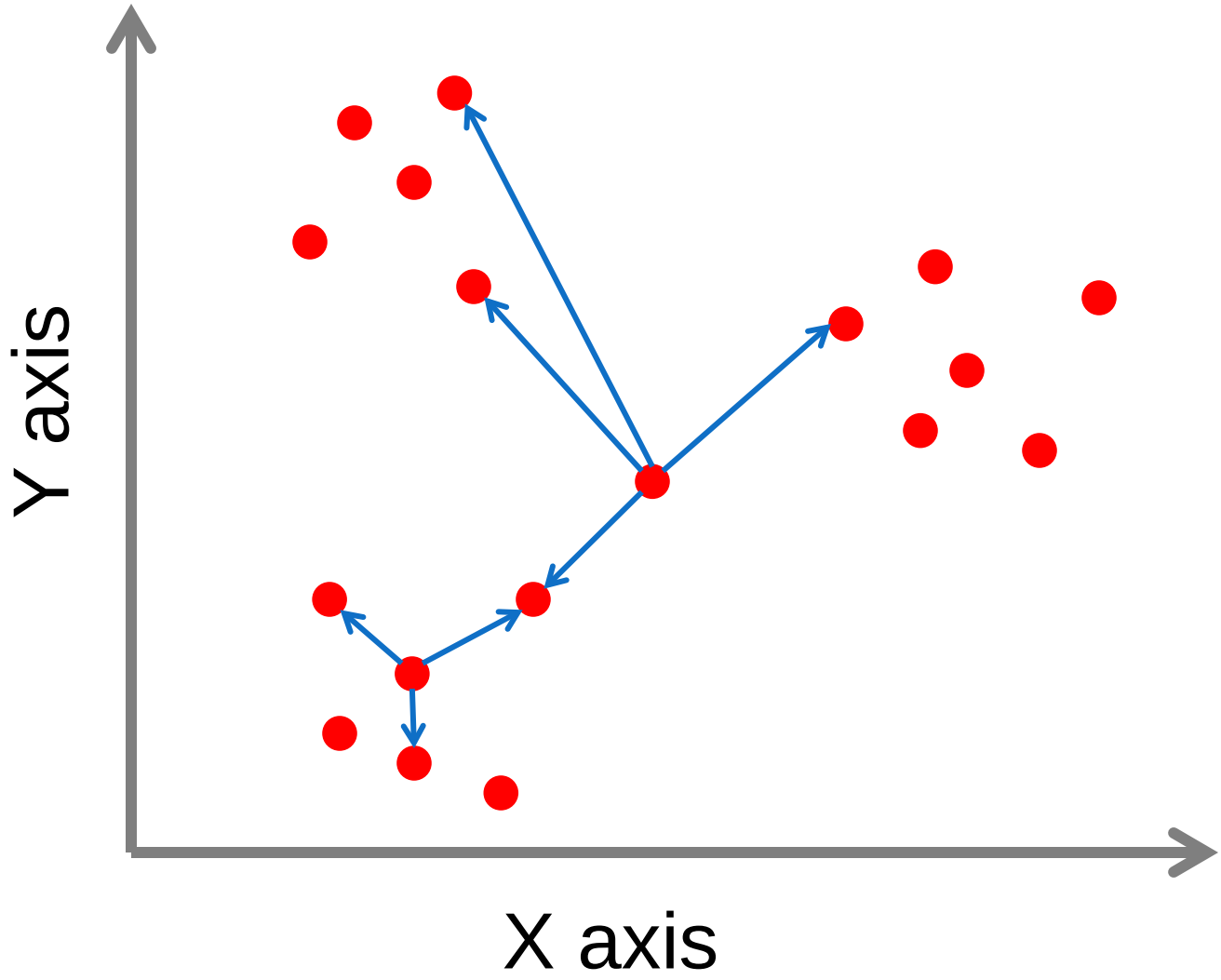
Euclidian Distance

- Straight line distance between two points

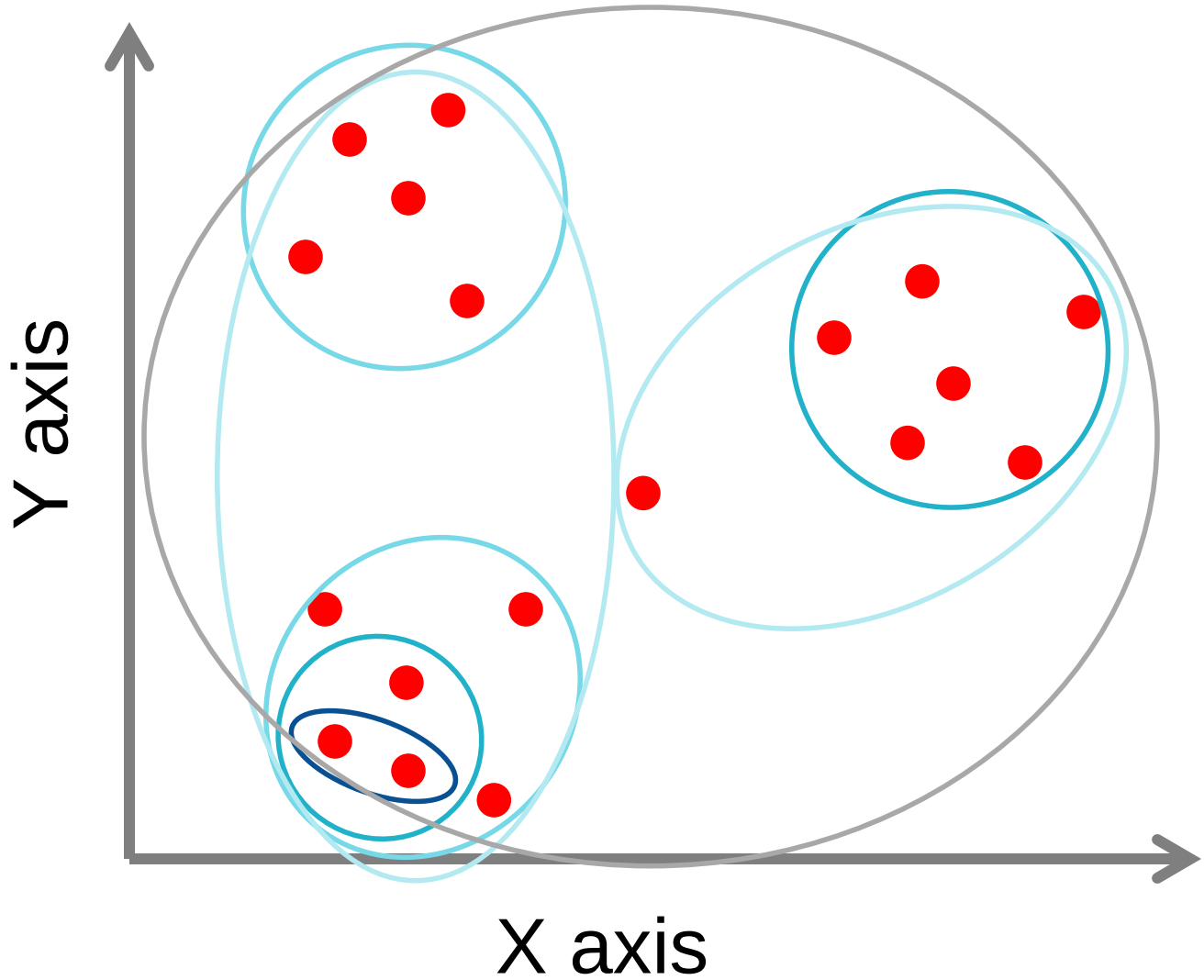
Euclidian Distance



Euclidian Distance



Clustering

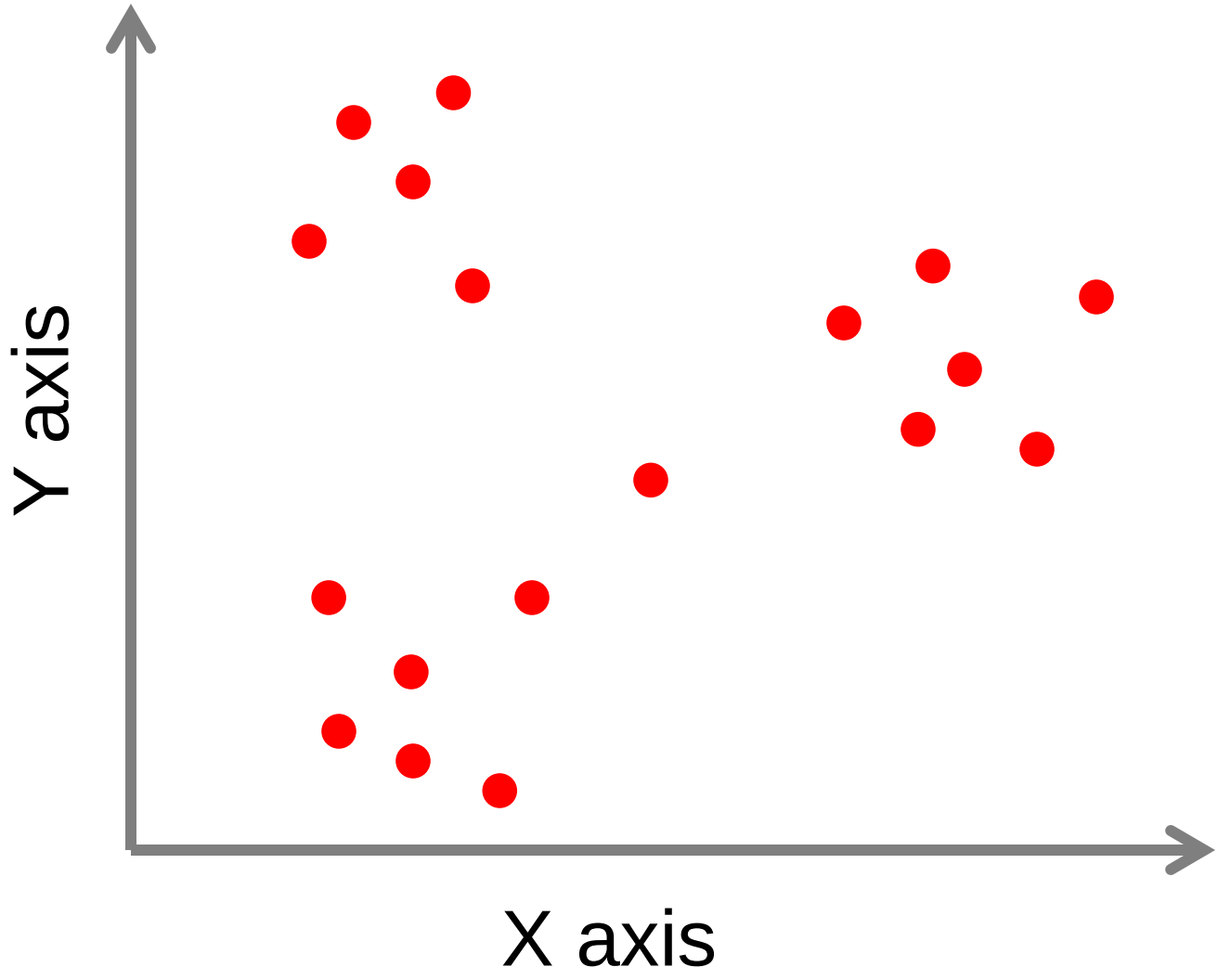


GENERAL ASSUMPTIONS

General Assumptions

- Components (X & Y axis) are uncorrelated
- Some relationship among variables
- On a graph, points that are closer together share more similarities than points that are farther apart
- Large variances have more important dynamics in defining clusters
- Data is normalized/standardized
  - Euclidean Distance (straight line distance between 2 points) assumes all parameters have the same scale for fair comparison between them

General Assumptions



Normalization/Standardization

Gene	cell 1	cell 2	cell 3	cell 4	cell 5	cell 6	cell 7	cell 8	cell 9	cell 10
a	12	8	12	8	20	8	8	20	8	24
Size	4789	2334	1566	4678	2346	9654	2345	3567	1245	2366
Grow	0.2	0.05	0.08	0.13	0.67	0.23	0.05	0.76	0.08	0.23
# MITO	20	20	20	20	8	20	20	8	24	8

- **Normalization → scales all numeric variables in the range [0,1]**

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- **Standardization → transforms data to have zero mean and unit variance [-1,+1]**

$$x_{new} = \frac{x - \mu}{\sigma}$$

Cell Example Raw Data

Gene	cell 1	cell 2	cell 3	cell 4	cell 5	cell 6	cell 7	cell 8	cell 9	cell 10
a	12	8	12	8	20	8	8	20	8	24
b	28	28	28	28	0	8	16	12	20	16
c	16	16	16	12	16	16	16	16	16	12
d	20	20	20	20	8	20	20	8	24	8
e	28	24	24	24	4	12	8	20	12	8
f	4	32	12	12	28	0	8	16	16	4
g	18	12	18	12	30	12	12	30	12	36
h	42	42	42	42	0	12	24	18	30	24
i	24	24	24	18	24	24	24	24	24	18
j	30	30	30	30	12	30	30	12	36	12
k	8	12	8	8	20	0	4	28	4	20
l	7	7	7	7	0	2	4	3	5	4
m	8	8	8	6	8	8	8	8	8	6
n	15	15	15	15	12	30	30	12	36	12
o	21	21	21	21	0	6	12	9	15	12

Cell Example PCA Results

Cells	Factor 1	Factor 2
cell1	0.93549	0.07226
cell2	0.81307	-0.0181
cell3	0.96315	0.0835
cell4	0.95191	-0.0752
cell5	-0.33566	0.75692
cell6	0.60245	0.0404
cell7	0.8073	0.0129
cell8	-0.01531	0.95305
cell9	0.8369	-0.07732
cell10	0.18234	0.85819

PCA and Cluster Analysis

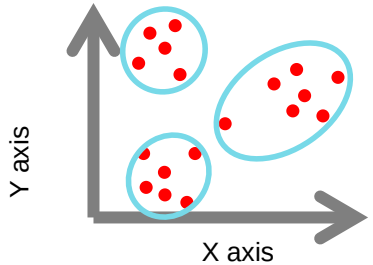
Cell Example

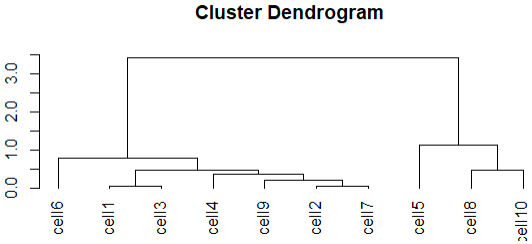
Gene	cell 1	cell 2	cell 3	cell 4	cell 5	cell 6	cell 7	cell 8	cell 9	cell 10
a	12	8	12	8	20	8	8	20	8	24
b	28	28	28	28	0	8	16	12	20	16
c	16	16	16	12	16	16	16	16	16	12
d	20	20	20	20	8	20	20	8	24	8
e	28	24	24	24	4	12	8	20	12	8
f	4	32	12	12	28	0	8	16	16	4
g	18	12	18	12	30	12	12	30	12	36
h	42	42	42	42	0	12	24	18	30	24
i	24	24	24	18	24	24	24	24	24	18
j	30	30	30	30	12	30	30	12	36	12
k	8	12	8	8	20	0	4	28	4	20
l	7	7	7	7	0	2	4	3	5	4
m	8	8	8	6	8	8	8	8	8	6
n	15	15	15	15	12	30	30	12	36	12
o	21	21	21	21	0	6	12	9	15	12

PCA

Cells	Factor 1	Factor 2
cell1	0.93549	0.07226
cell2	0.81307	-0.0181
cell3	0.96315	0.0835
cell4	0.95191	-0.0752
cell5	-0.33566	0.75692
cell6	0.60245	0.0404
cell7	0.8073	0.0129
cell8	-0.01531	0.95305
cell9	0.8369	-0.07732
cell10	0.18234	0.85819

Cluster Analysis



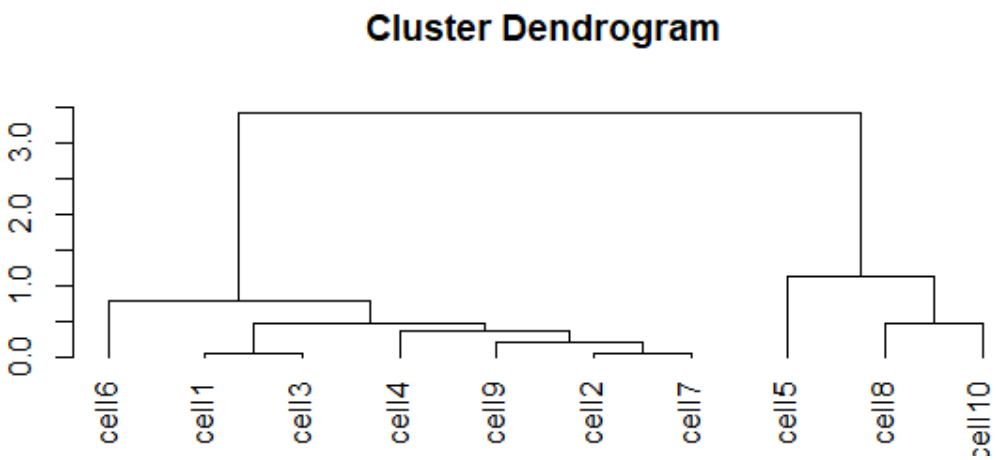


HIERARCHICAL CLUSTER

How many clusters?

Hierarchical Cluster

- A series of steps that build a tree-like structure by either adding elements (i.e. agglomerative) to form a large cluster or by subtracting elements (i.e. divisive) from a large cluster to form smaller clusters
- Dendogram is used to visualize the results



Single Linkage

Complete Linkage

Average Linkage

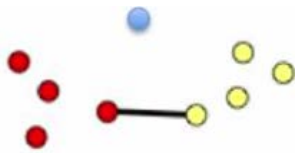
Centroid Linkage

Ward's Linkage

Agglomerative Clustering

Single Linkage

$$D(c_1, c_2) = \min_{x_1 \in c_1, x_2 \in c_2} D(x_1, x_2)$$



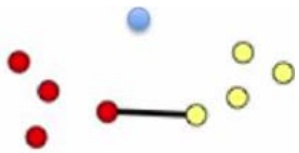
- Distance between closest elements in cluster
- Produces long chains a → b → c → … → z

D (distance); c1, c2 (clusters); x1, y2 (distance between two elements) http://bit.ly/s-link

Single Linkage

$$D(c_1, c_2) = \min_{x_1 \in c_1, x_2 \in c_2} D(x_1, x_2)$$

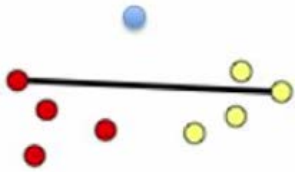
- Distance between closest elements in cluster
- Produces long chains a → b → c → … → z



Complete Linkage

$$D(c_1, c_2) = \max_{x_1 \in c_1, x_2 \in c_2} D(x_1, x_2)$$

- Distance between farthest elements in clusters
- Forces “spherical” clusters with consistent diameter

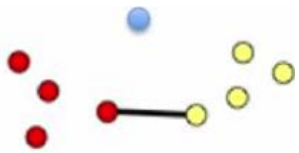


D (distance); c1, c2 (clusters); x1, y2 (distance between two elements)

Single Linkage

$$D(c_1,c_2) = \min_{x_1 \in c_1, x_2 \in c_2} D(x_1,x_2)$$

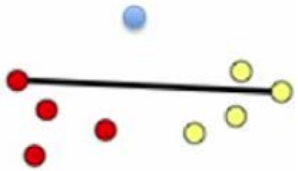
- Distance between closest elements in cluster
- Produces long chains a → b → c → … → z



Complete Linkage

$$D(c_1,c_2) = \max_{x_1 \in c_1, x_2 \in c_2} D(x_1,x_2)$$

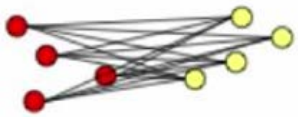
- Distance between farthest elements in clusters
- Forces “spherical” clusters with consistent diameter



Average Linkage

$$D(c_1,c_2) = \frac{1}{|c_1|}\frac{1}{|c_2|}\sum_{x_1 \in c_1}\sum_{x_2 \in c_2} D(x_1,x_2)$$

- Average of all pairwise distances
- Less affected by outliers

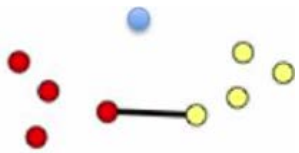


D (distance); c1, c2 (clusters); x1, y2 (distance between two elements)

Single Linkage

$$D(c_1,c_2) = \min_{x_1 \in c_1, x_2 \in c_2} D(x_1,x_2)$$

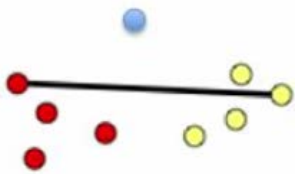
- Distance between closest elements in cluster
- Produces long chains a → b → c → … → z



Complete Linkage

$$D(c_1,c_2) = \max_{x_1 \in c_1, x_2 \in c_2} D(x_1,x_2)$$

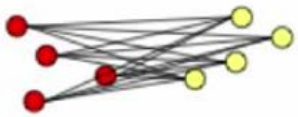
- Distance between farthest elements in clusters
- Forces “spherical” clusters with consistent diameter



Average Linkage

$$D(c_1,c_2) = \frac{1}{|c_1|}\frac{1}{|c_2|}\sum_{x_1 \in c_1}\sum_{x_2 \in c_2} D(x_1,x_2)$$

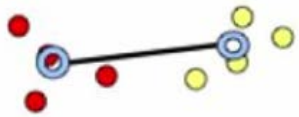
- Average of all pairwise distances
- Less affected by outliers



Centroid Linkage

$$D(c_1,c_2) = D\left(\left(\frac{1}{|c_1|}\sum_{x \in c_1}\vec{x}\right),\left(\frac{1}{|c_2|}\sum_{x \in c_2}\vec{x}\right)\right)$$

- Distance between centroids (means) of two clusters
- Requires → numerical data

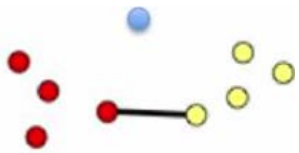


D (distance); c1, c2 (clusters); x1, y2 (distance between two elements)

Single Linkage

$$D(c_1,c_2) = \min_{x_1 \in c_1, x_2 \in c_2} D(x_1,x_2)$$

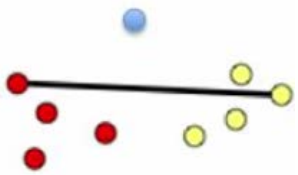
- Distance between closest elements in cluster
- Produces long chains a → b → c → … → z



Complete Linkage

$$D(c_1,c_2) = \max_{x_1 \in c_1, x_2 \in c_2} D(x_1,x_2)$$

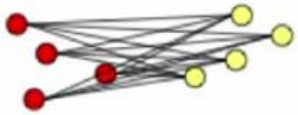
- Distance between farthest elements in clusters
- Forces "spherical" clusters with consistent diameter



Average Linkage

$$D(c_1,c_2) = \frac{1}{|c_1|}\frac{1}{|c_2|}\sum_{x_1 \in c_1}\sum_{x_2 \in c_2} D(x_1,x_2)$$

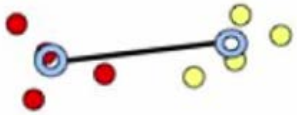
- Average of all pairwise distances
- Less affected by outliers



Centroid Linkage

$$D(c_1,c_2) = D\left(\left(\frac{1}{|c_1|}\sum_{x \in c_1}\vec{x}\right),\left(\frac{1}{|c_2|}\sum_{x \in c_2}\vec{x}\right)\right)$$

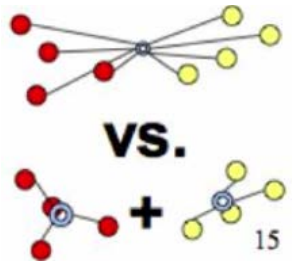
- Distance between centroids (means) of two clusters
- Requires → numerical data



Ward's Linkage

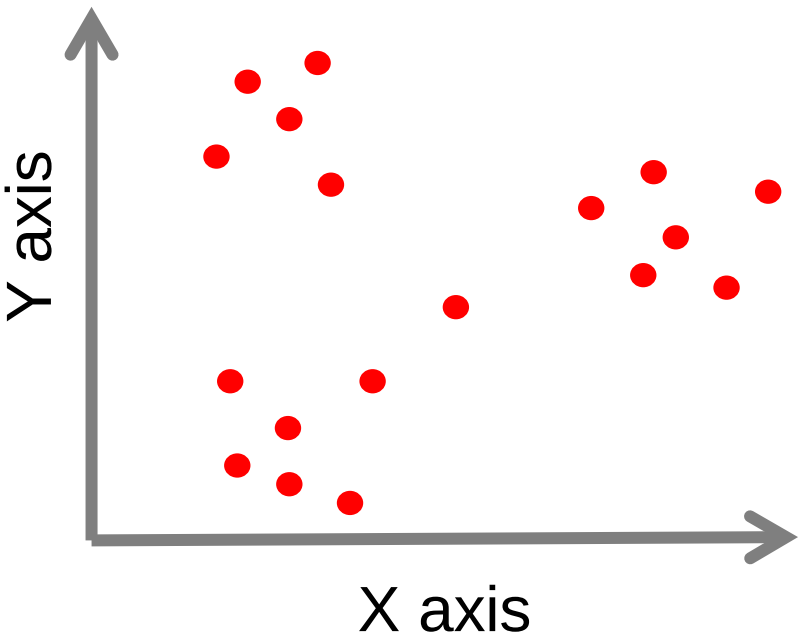
$$TD_{c_1 \cup c_2} = \sum_{x \in c_1 \cup c_2} D(x, \mu_{c_1 \cup c_2})^2$$

- Consider joining two clusters
- Requires → numerical data



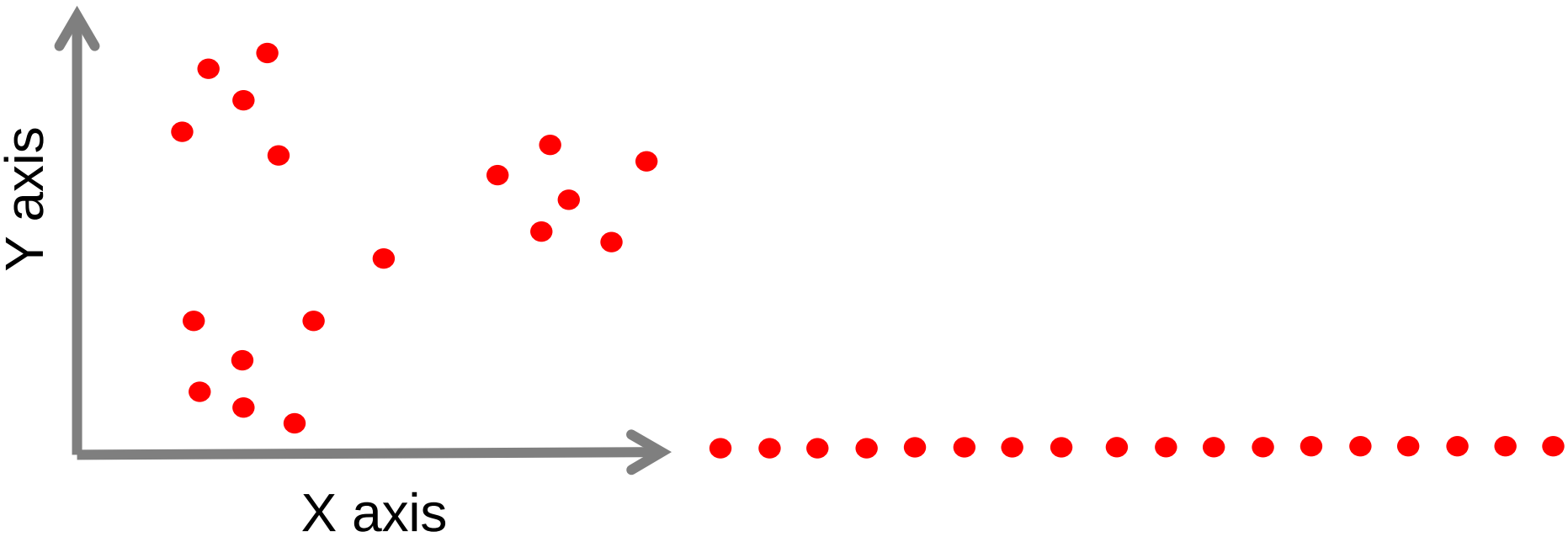
D (distance); c1, c2 (clusters); x1, y2 (distance between two elements)

Single Linkage Example

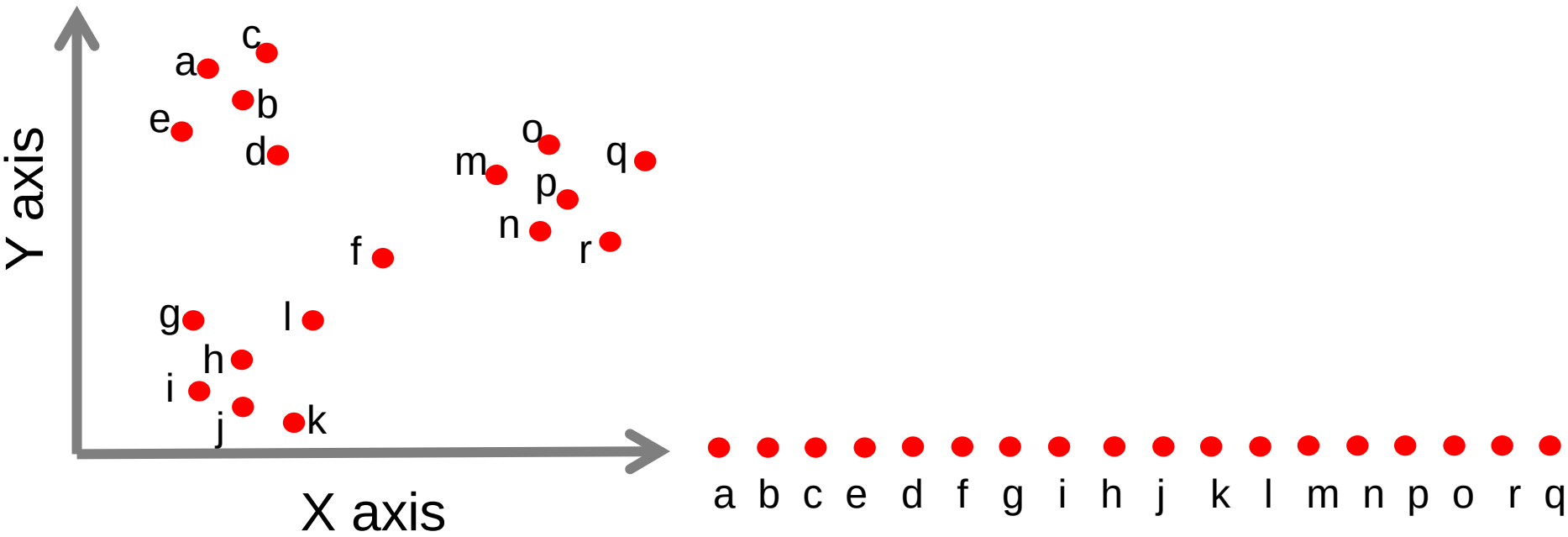


Item	X	Y
a		
b		
c		
d		
e		
f		
g		
h		
i		
j		
Etc.		

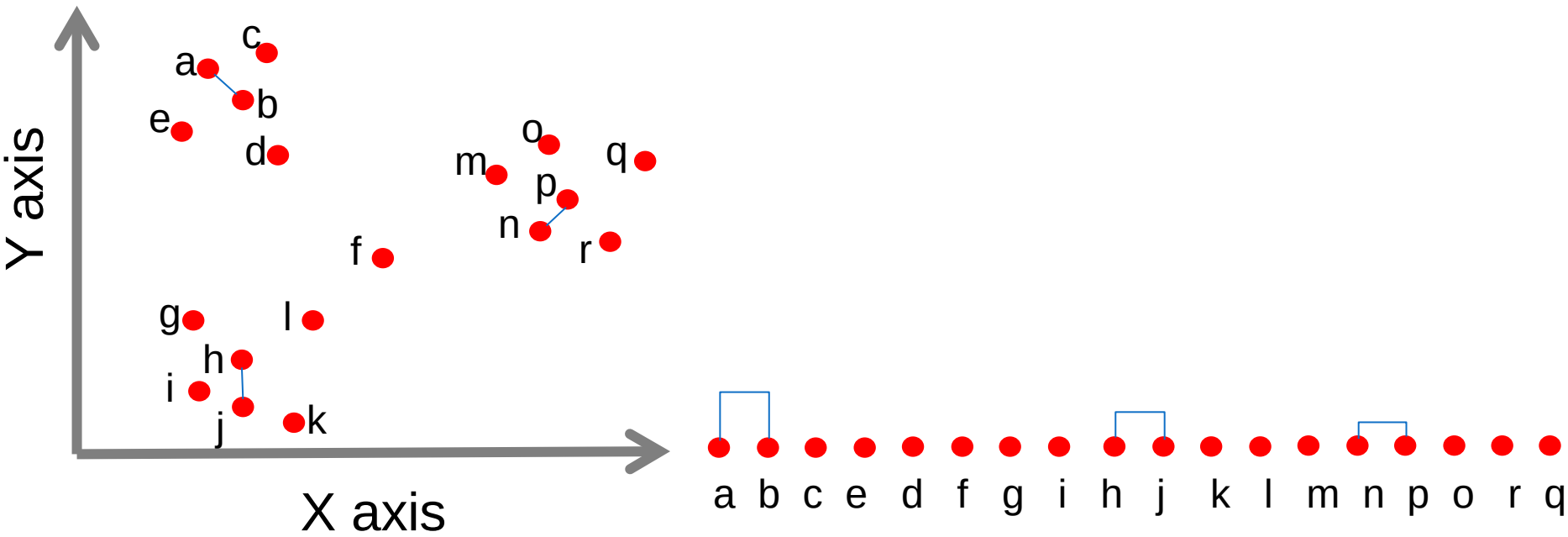
Single Linkage Example



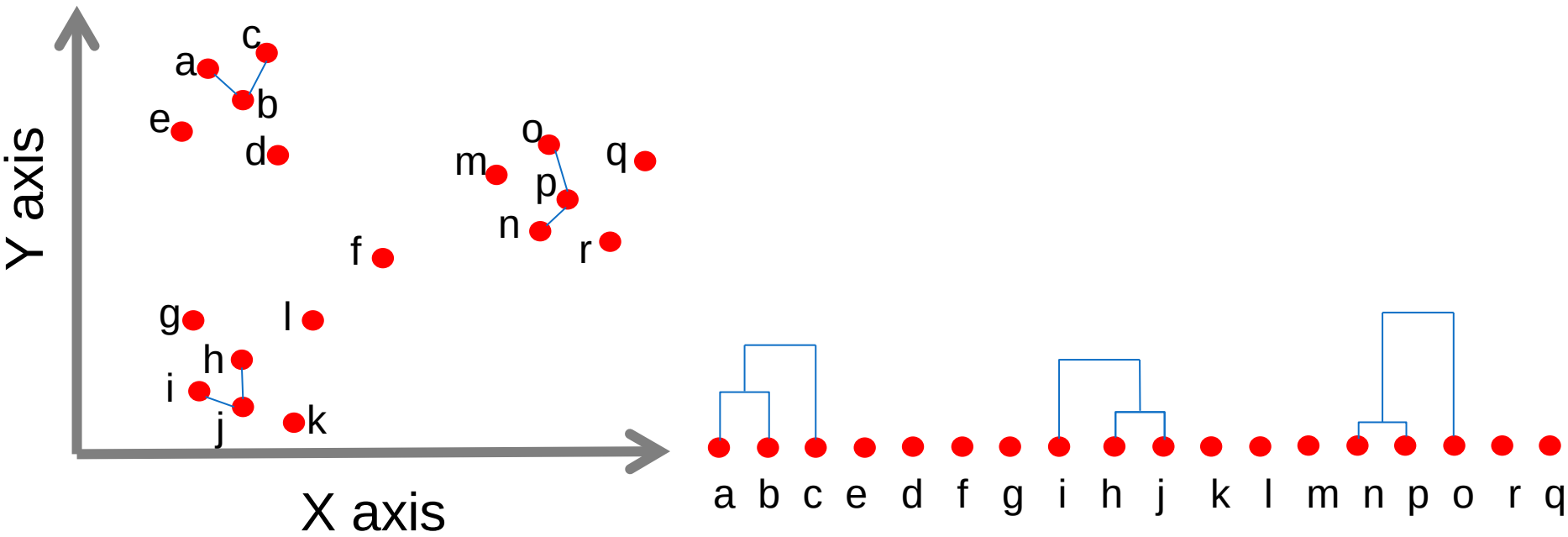
Single Linkage Example



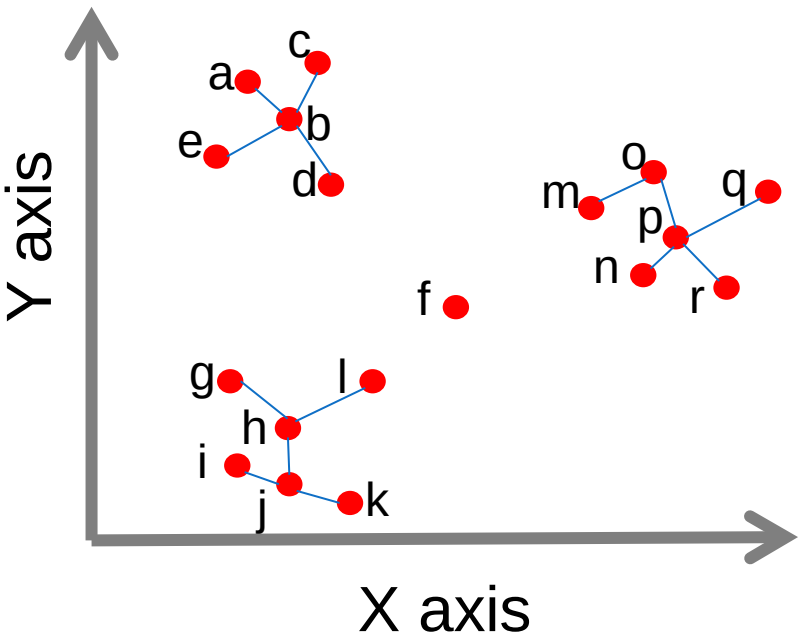
Single Linkage Example

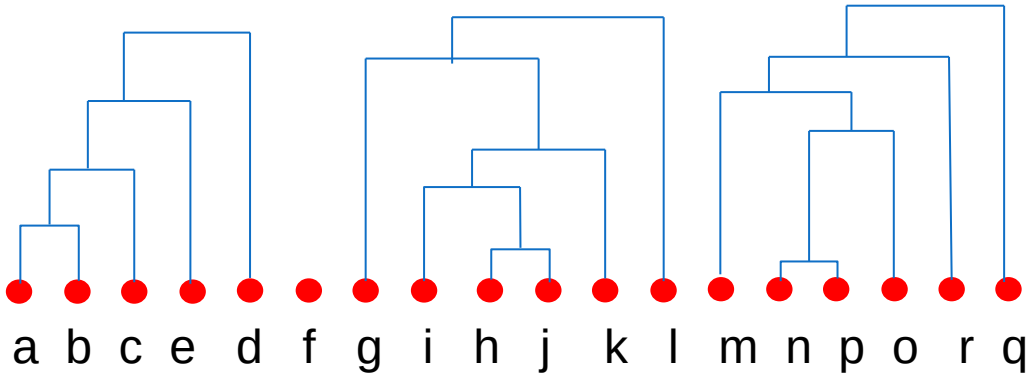


Single Linkage Example

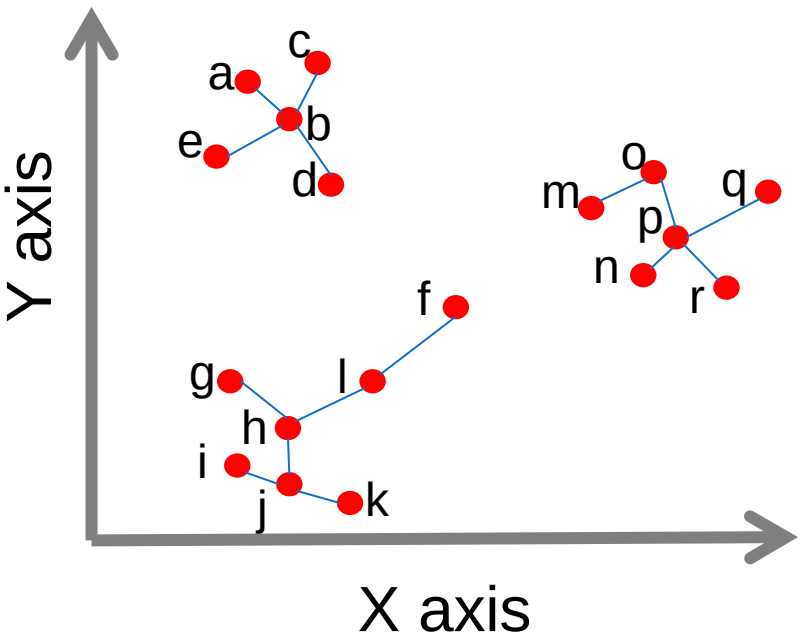


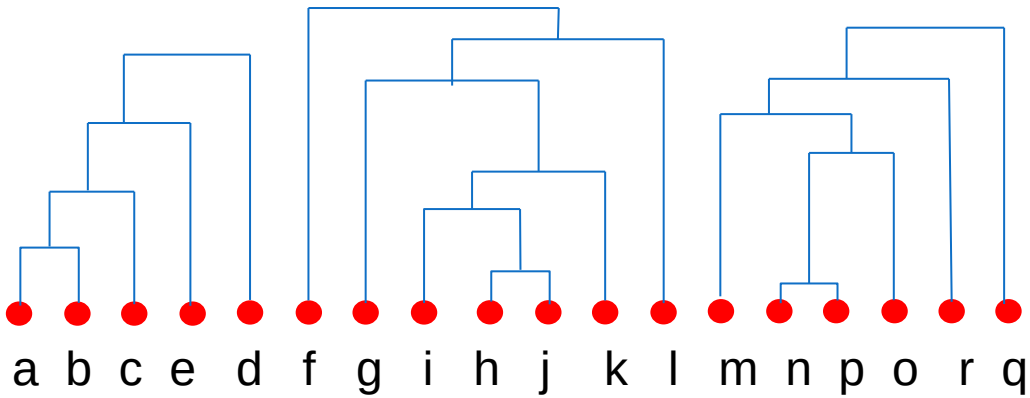
Single Linkage Example



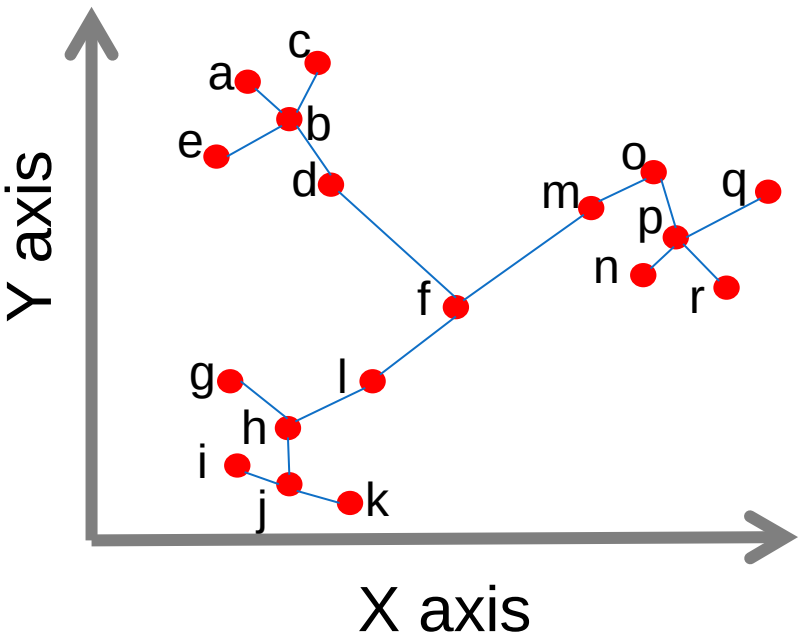


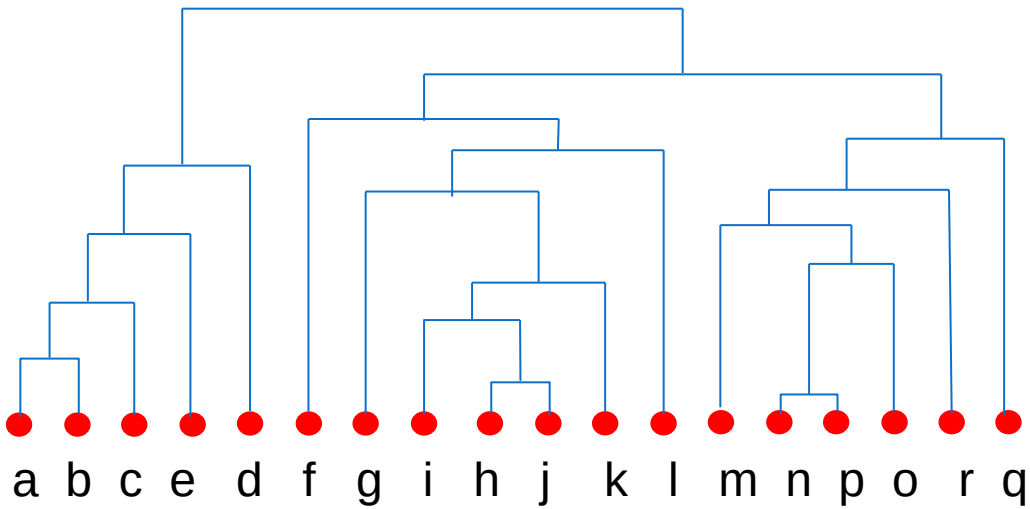
Single Linkage Example





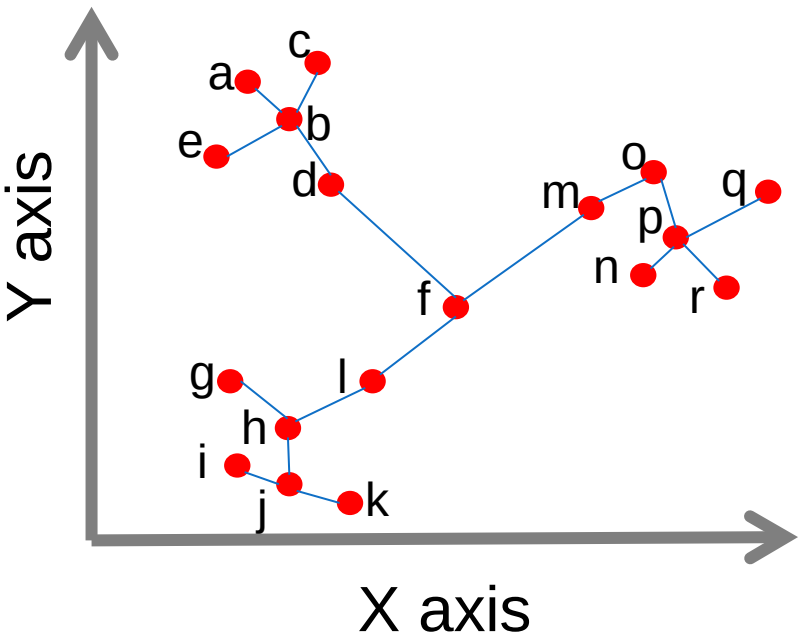
Single Linkage Example

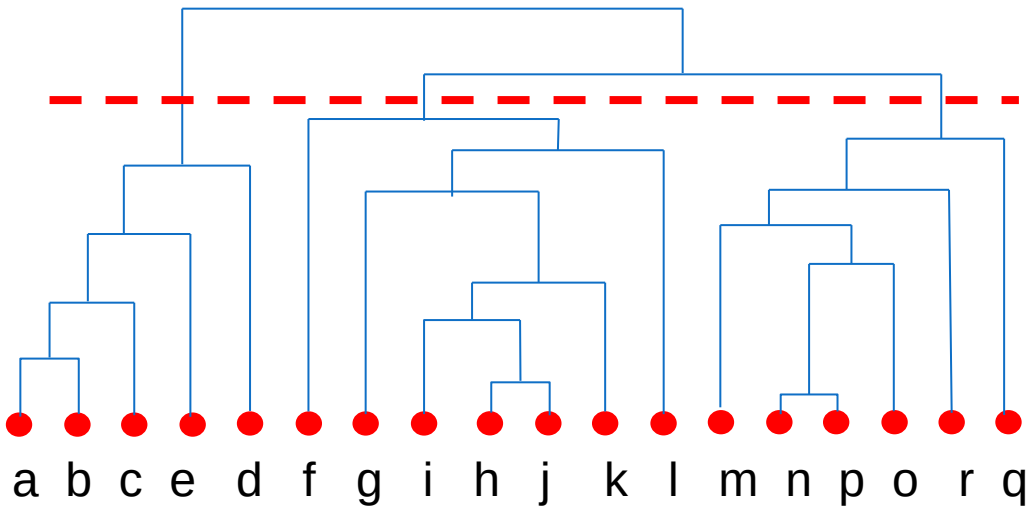




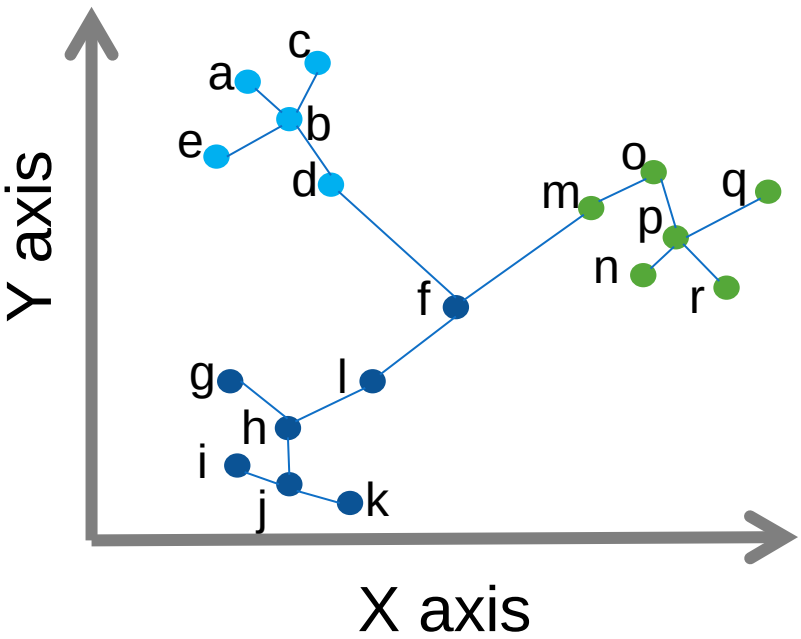
How many clusters do we have?

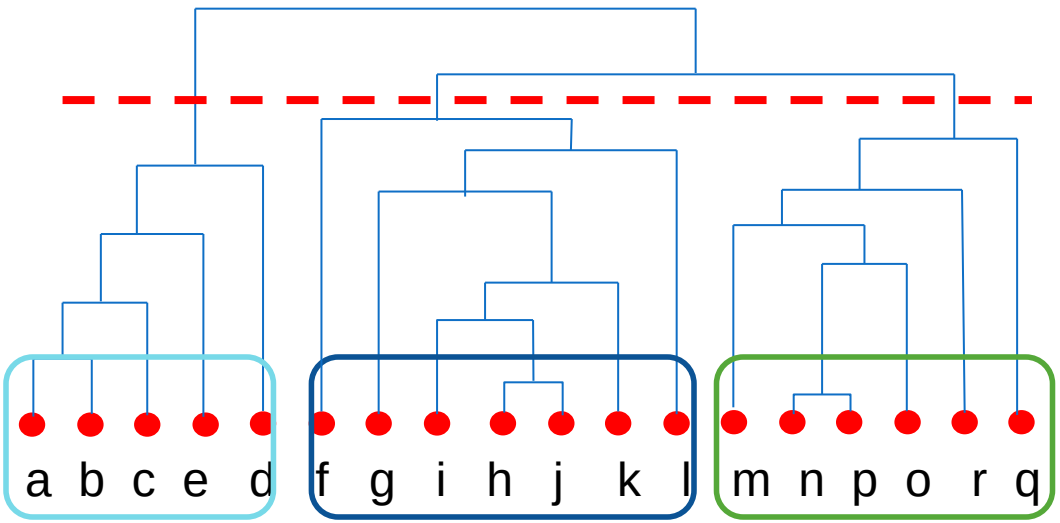
Single Linkage Example



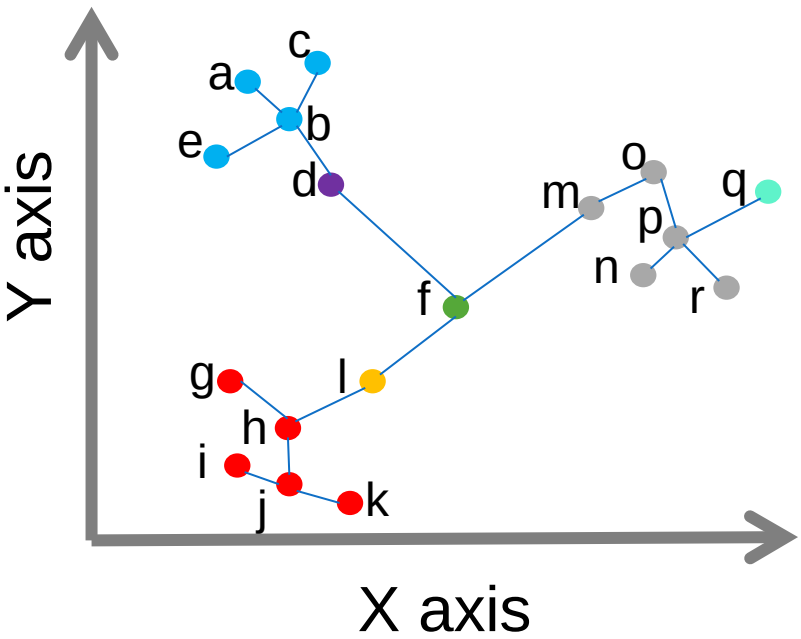


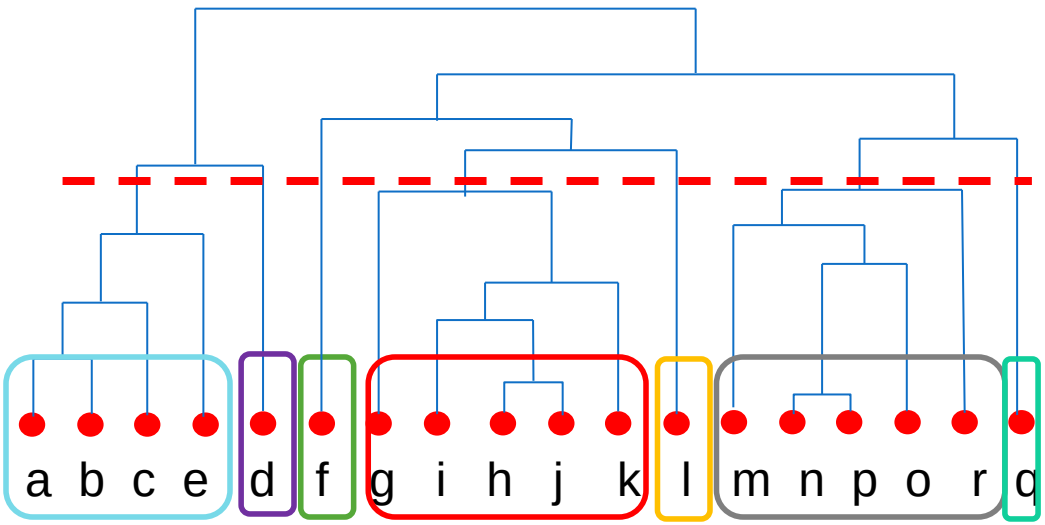
Single Linkage Example





Single Linkage Example





Limitations of Hierarchical Clustering

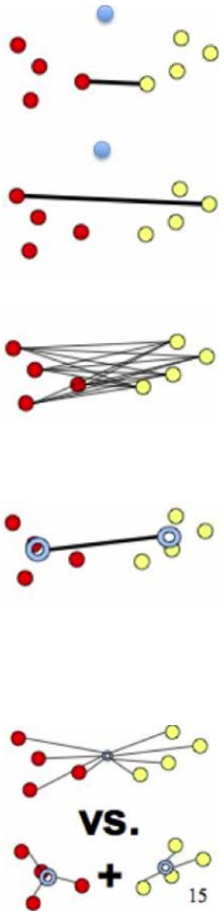
- Single, Complete and Average Linkage can use numerical or categorical data as long as the distance is defined
- Centroid and Ward's Linkage requires numerical (i.e. interval or ratio) data since the formula uses means

Limitations of Hierarchical Clustering

- Underlying structure of the sample is unknown which makes it difficult to select the “correct” algorithm
- Poor cluster assignments cannot be modified
- Unstable solutions with a small sample (need at least 150 observations)

Limitations of Hierarchical Clustering

- Outliers can affect clustering
  - Single and Complete Linkage → outliers can merge the wrong clusters
  - Average Linkage → is less affected by outliers because it computes average distances
  - Centroid Linkage → produces irregular shaped clusters where outliers influence the position of the centroid
  - Ward's Linkage → tends to produce clusters with similar number of observations which makes it easy for outliers to distort results



HIERARCHICAL CLUSTER IN R

PCA Results

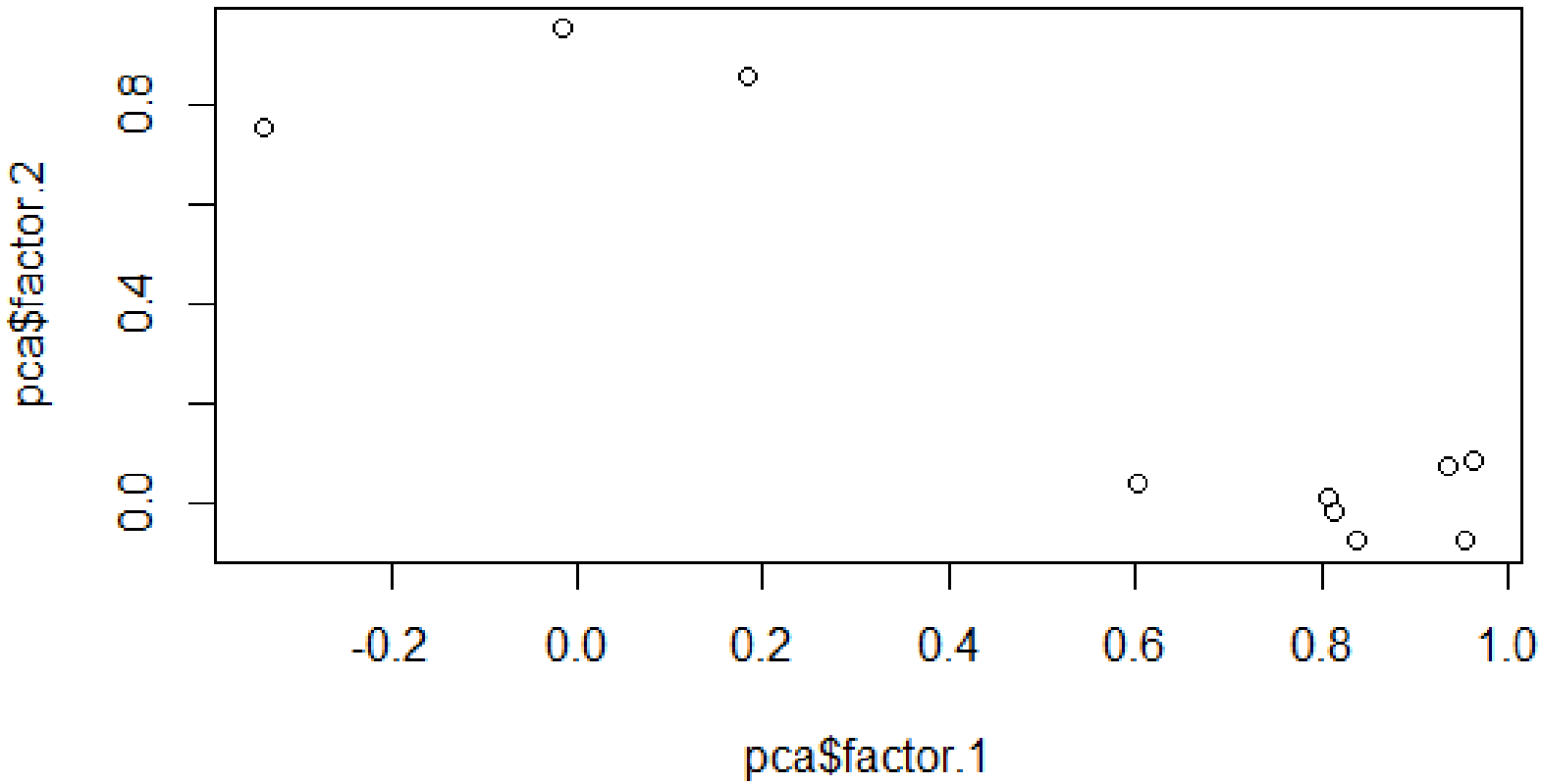
Cells	Factor 1	Factor 2
cell1	0.93549	0.07226
cell2	0.81307	-0.0181
cell3	0.96315	0.0835
cell4	0.95191	-0.0752
cell5	-0.33566	0.75692
cell6	0.60245	0.0404
cell7	0.8073	0.0129
cell8	-0.01531	0.95305
cell9	0.8369	-0.07732
cell10	0.18234	0.85819

Open File

```
> pca<-read.csv(file.choose(),header=TRUE)
> pca
    cells factor.1 factor.2
1   cell1  0.93549  0.07226
2   cell2  0.81307 -0.01810
3   cell3  0.96315  0.08350
4   cell4  0.95191 -0.07520
5   cell5 -0.33566  0.75692
6   cell6  0.60245  0.04040
7   cell7  0.80730  0.01290
8   cell8 -0.01531  0.95305
9   cell9  0.83690 -0.07732
10 cell10  0.18234  0.85819
```

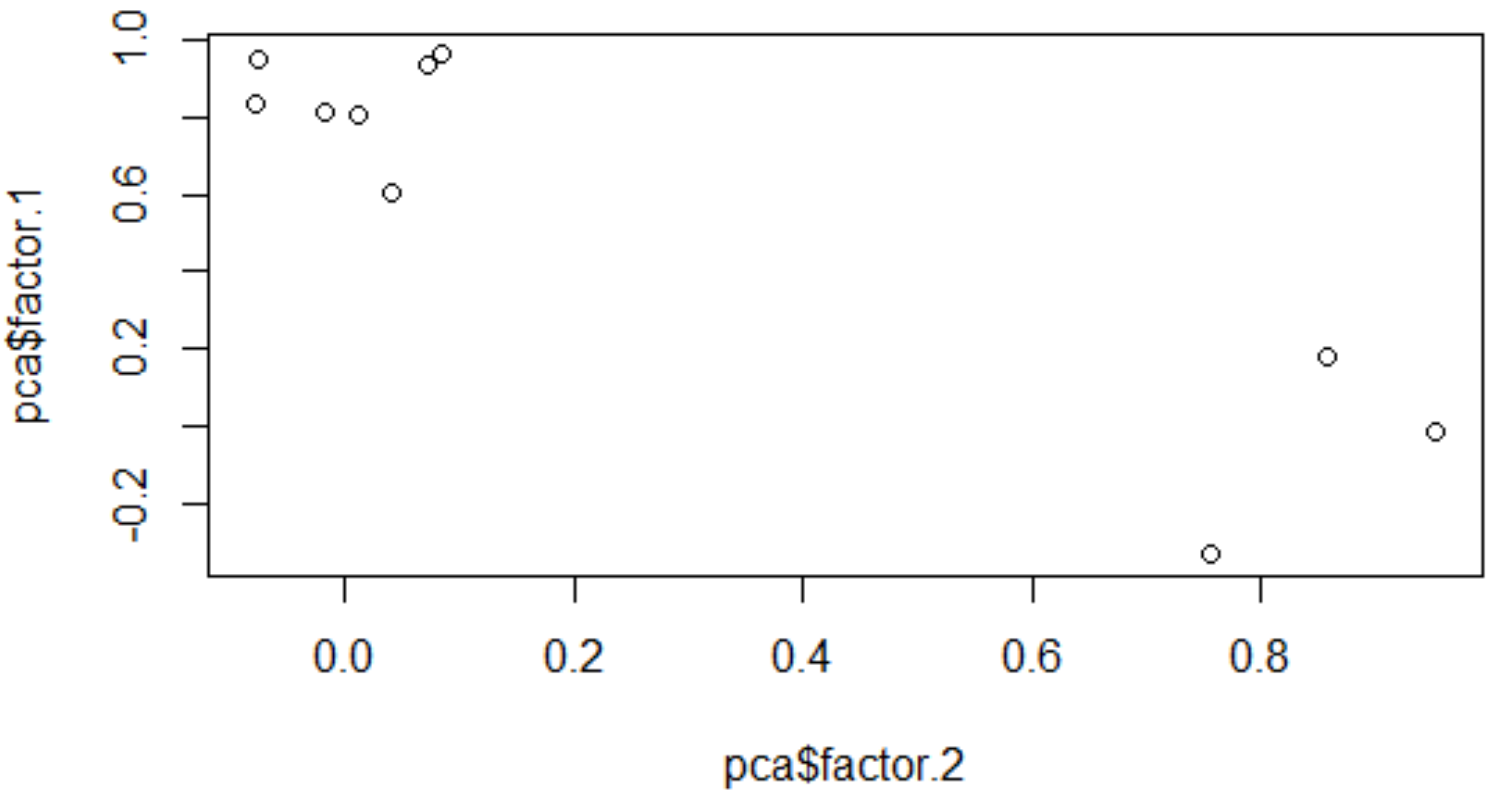
Graph Data

```
> plot(pca$factor.2~pca$factor.1,data=pca)
```



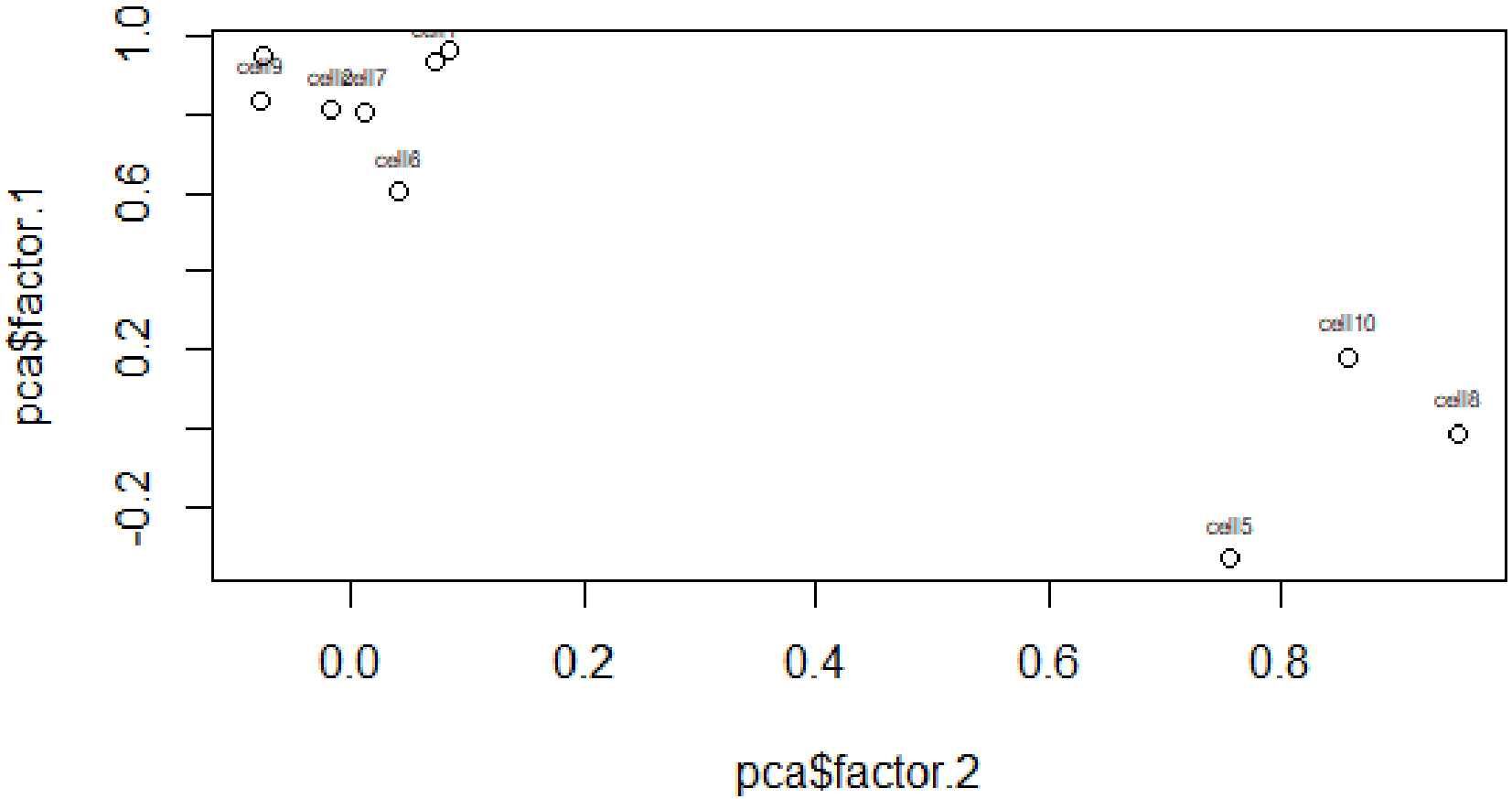
Graph Data

```
> plot(pca$factor.1~pca$factor.2,data=pca)
```



Graph Data

```
> with(pca,text(pca$factor.1~pca$factor.2, labels=pca$cells,pos=3.5, cex=.5))
```



Standardize Data

- Subtract first column to have quantitative data

```
> pca
    cells factor.1 factor.2
1   cell1  0.93549  0.07226
2   cell2  0.81307 -0.01810
3   cell3  0.96315  0.08350
4   cell4  0.95191 -0.07520
5   cell5 -0.33566  0.75692
6   cell6  0.60245  0.04040
7   cell7  0.80730  0.01290
8   cell8 -0.01531  0.95305
9   cell9  0.83690 -0.07732
10 cell10  0.18234  0.85819
```

```
> z=pca[,-c(1,1)]
> z
   factor.1 factor.2
1   0.93549  0.07226
2   0.81307 -0.01810
3   0.96315  0.08350
4   0.95191 -0.07520
5  -0.33566  0.75692
6   0.60245  0.04040
7   0.80730  0.01290
8  -0.01531  0.95305
9   0.83690 -0.07732
10  0.18234  0.85819
```

Standardize

- Subtract mean and divide by standard deviation

```
> m<-apply(z,2,mean)
> s<-apply(z,2,sd)
> z<-scale(z,m,s)
> z
         factor.1   factor.2
 [1,]  0.77933569 -0.4519086
 [2,]  0.51529083 -0.6686520
 [3,]  0.83899491 -0.4249476
 [4,]  0.81475161 -0.8056157
 [5,] -1.96237834  1.1903617
 [6,]  0.06100942 -0.5283300
 [7,]  0.50284565 -0.5942933
 [8,] -1.27142284  1.6608119
 [9,]  0.56668920 -0.8107009
[10,] -0.84511612  1.4332745
attr(,"scaled:center")
factor.1 factor.2
0.574164 0.260660
attr(,"scaled:scale")
 factor.1  factor.2
0.4636333 0.4168985
>
```

```
> z=pca[,-c(1,1)]
> z
   factor.1 factor.2
1   0.93549  0.07226
2   0.81307 -0.01810
3   0.96315  0.08350
4   0.95191 -0.07520
5  -0.33566  0.75692
6   0.60245  0.04040
7   0.80730  0.01290
8  -0.01531  0.95305
9   0.83690 -0.07732
10  0.18234  0.85819
```

Euclidian Distance

- Measures the distance between all the points

```
> distance<-dist(z)
> distance
           1          2          3          4          5          6          7          8          9
2  0.34161000
3  0.06546845 0.40518658
4  0.35547583 0.32929598 0.38143939
5  3.19594232 3.09754357 3.23371551 3.41999658
6  0.72238001 0.47545961 0.78482441 0.80312805 2.65480681
7  0.31099867 0.07539289 0.37639649 0.37675256 3.04340644 0.44673304
8  2.94435023 2.93577036 2.96719659 3.23038530 0.83590845 2.56275597 2.86941253
9  0.41707372 0.15106191 0.47218223 0.24811453 3.22497041 0.57917645 0.22562855 3.08010251
10 2.48852544 2.50375756 2.50783157 2.78707561 1.14336412 2.16077668 2.43475507 0.48322943 2.65115446
>
```

```
> print(distance,digits=3)
       1      2      3      4      5      6      7      8      9
2  0.3416
3  0.0655 0.4052
4  0.3555 0.3293 0.3814
5  3.1959 3.0975 3.2337 3.4200
6  0.7224 0.4755 0.7848 0.8031 2.6548
7  0.3110 0.0754 0.3764 0.3768 3.0434 0.4467
8  2.9444 2.9358 2.9672 3.2304 0.8359 2.5628 2.8694
9  0.4171 0.1511 0.4722 0.2481 3.2250 0.5792 0.2256 3.0801
10 2.4885 2.5038 2.5078 2.7871 1.1434 2.1608 2.4348 0.4832 2.6512
```

Euclidian Distance

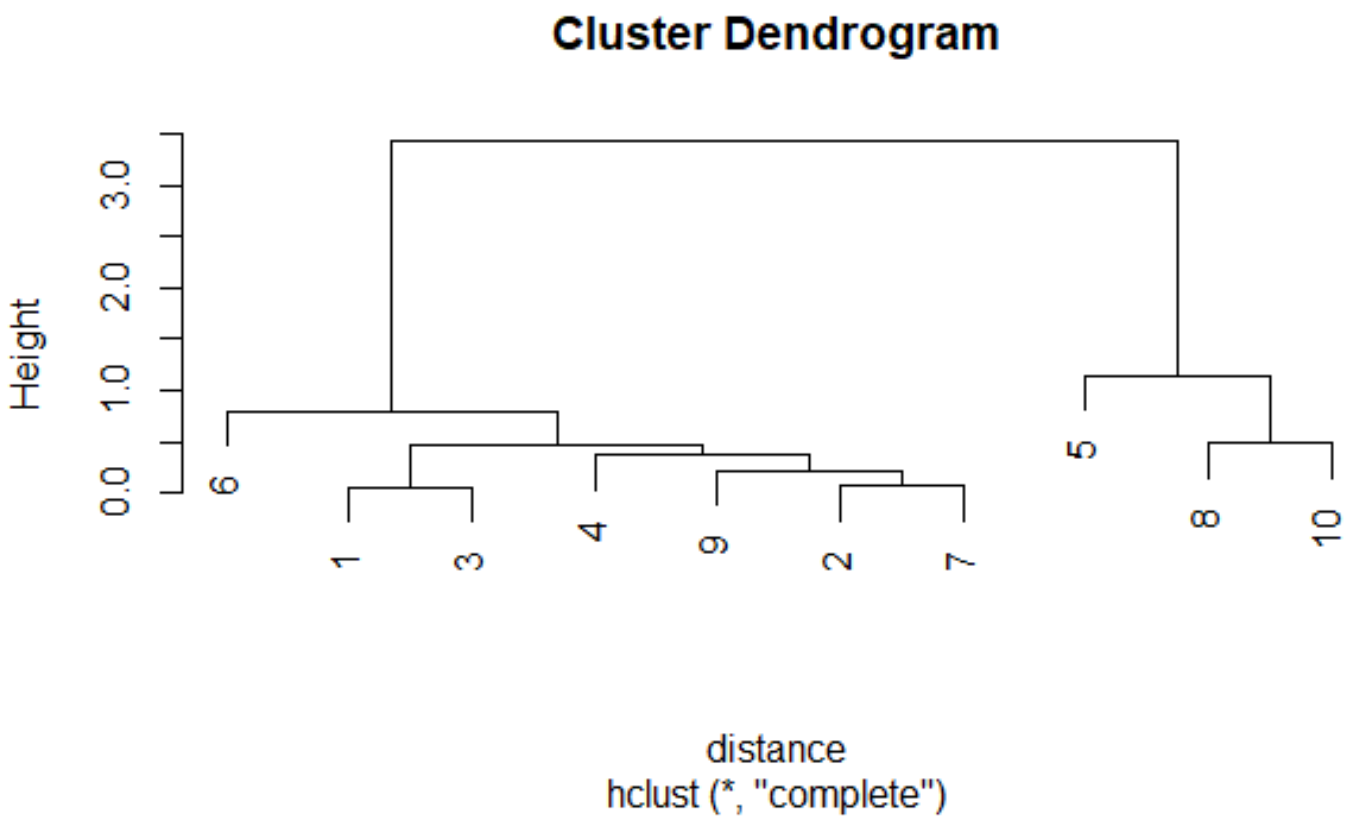
- Measures the distance between all the points

```
> distance<-dist(z)
> distance
           1          2          3          4          5          6          7          8          9
2  0.34161000
3  0.06546845 0.40518658
4  0.35547583 0.32929598 0.38143939
5  3.19594232 3.09754357 3.23371551 3.41999658
6  0.72238001 0.47545961 0.78482441 0.80312805 2.65480681
7  0.31099867 0.07539289 0.37639649 0.37675256 3.04340644 0.44673304
8  2.94435023 2.93577036 2.96719659 3.23038530 0.83590845 2.56275597 2.86941253
9  0.41707372 0.15106191 0.47218223 0.24811453 3.22497041 0.57917645 0.22562855 3.08010251
10 2.48852544 2.50375756 2.50783157 2.78707561 1.14336412 2.16077668 2.43475507 0.48322943 2.65115446
>
```

```
> print(distance,digits=3)
       1      2      3      4      5      6      7      8      9
2  0.3416
3  0.0655 0.4052
4  0.3555 0.3293 0.3814
5  3.1959 3.0975 3.2337 3.4200
6  0.7224 0.4755 0.7848 0.8031 2.6548
7  0.3110 0.0754 0.3764 0.3768 3.0434 0.4467
8  2.9444 2.9358 2.9672 3.2304 0.8359 2.5628 2.8694
9  0.4171 0.1511 0.4722 0.2481 3.2250 0.5792 0.2256 3.0801
10 2.4885 2.5038 2.5078 2.7871 1.1434 2.1608 2.4348 0.4832 2.6512
```

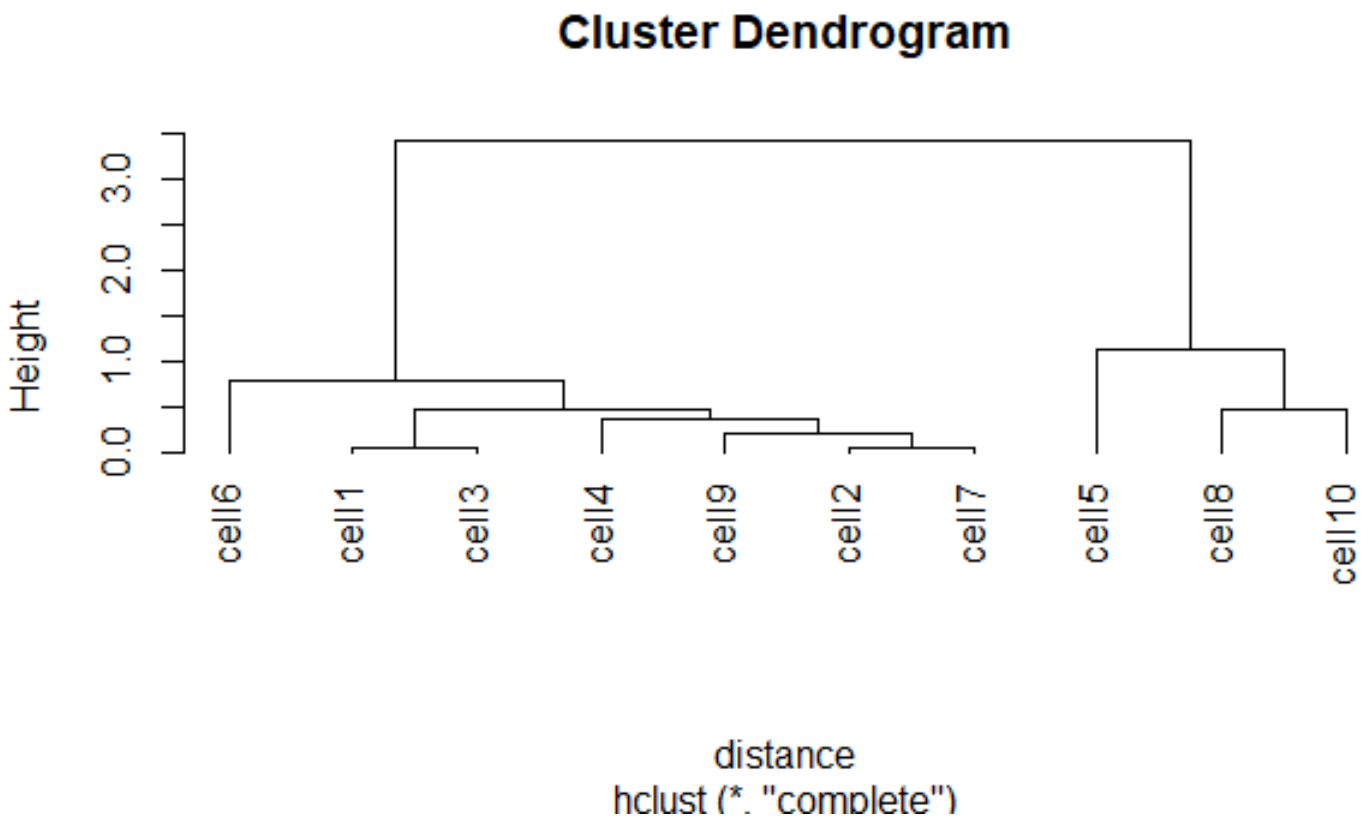
Hierarchical Clustering (complete)

```
> hc.c<-hclust(distance)
> plot(hc.c)
```

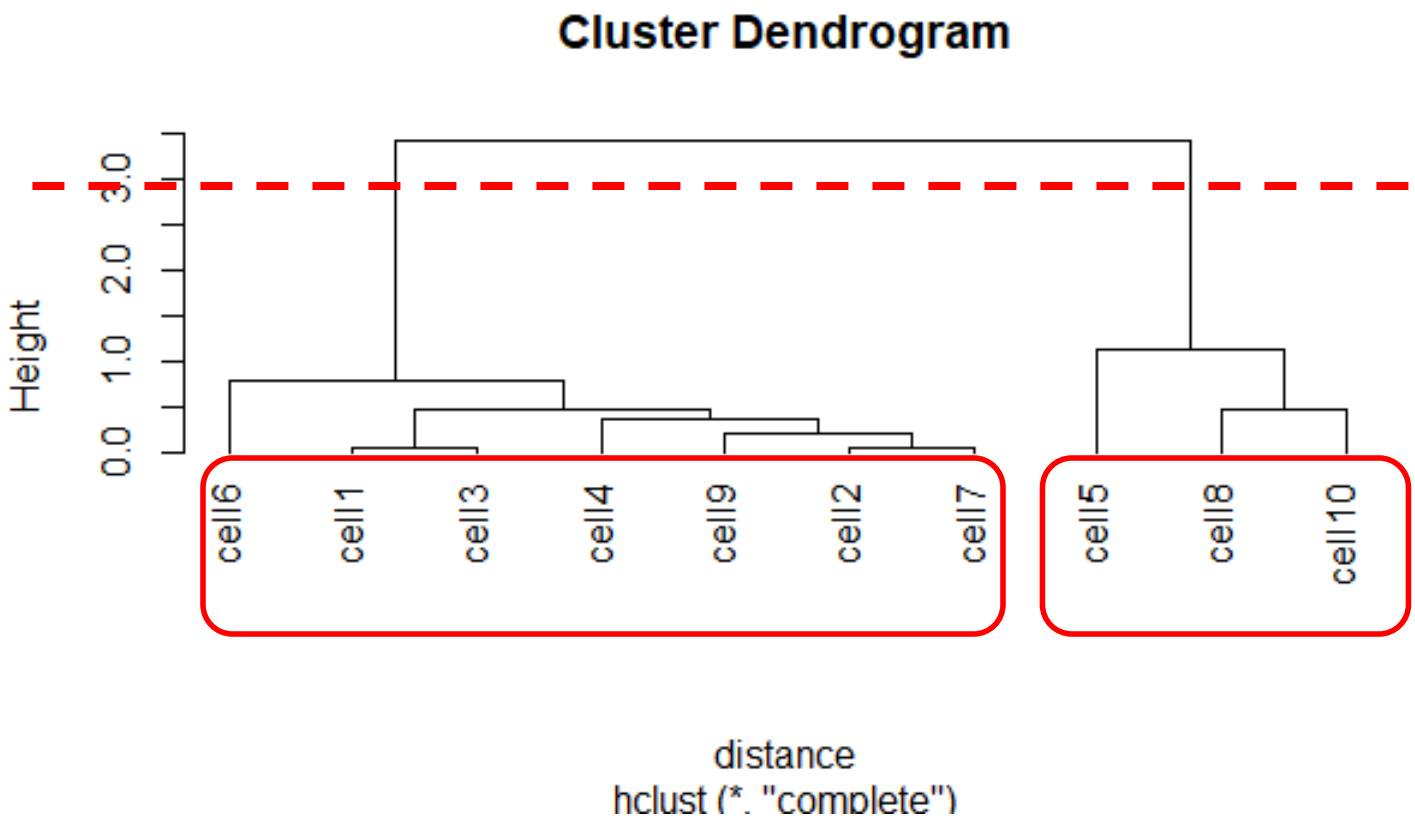


Hierarchical Clustering (complete)

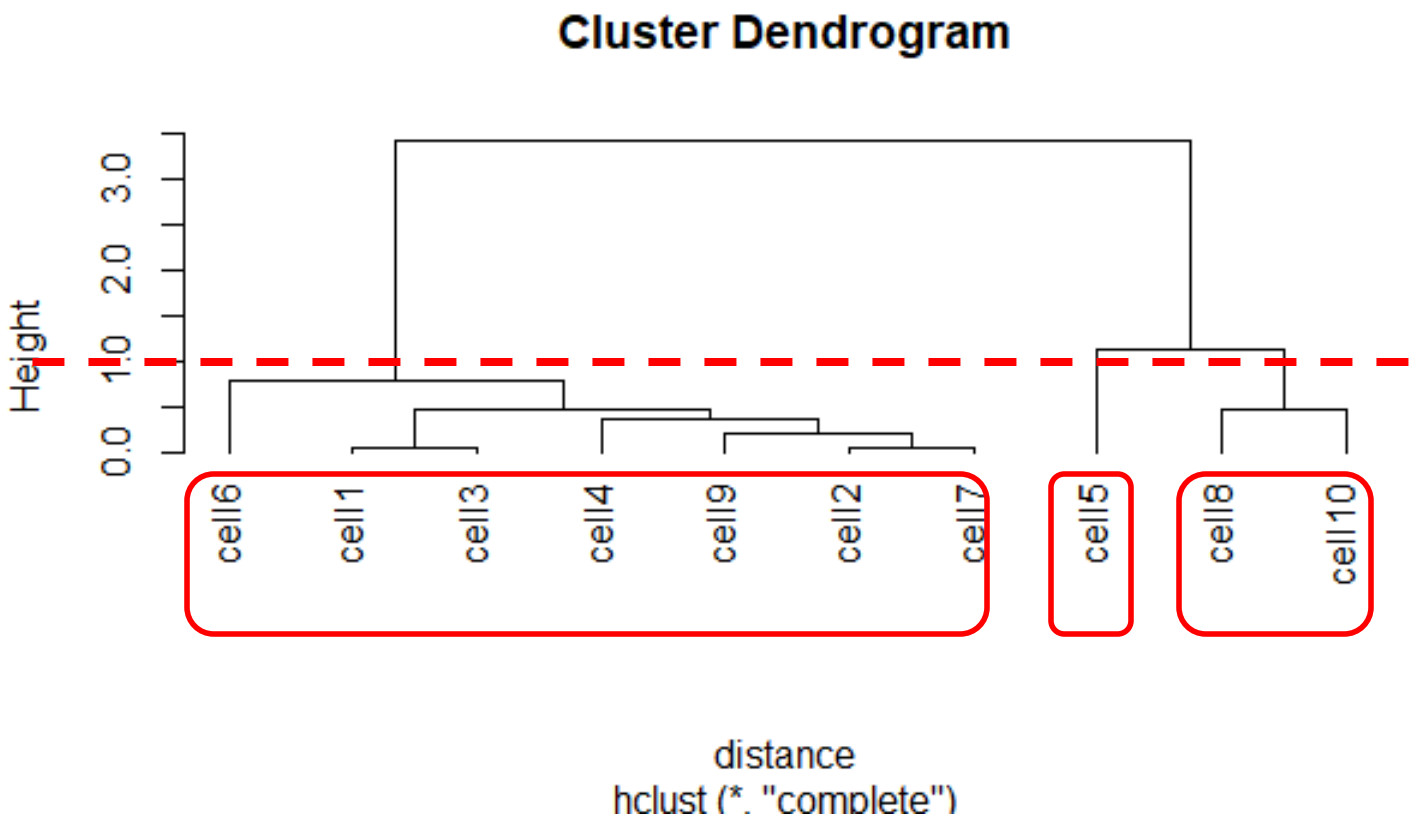
```
> plot(hc.c,hang=-1,labels=pca$cells)
```



Hierarchical Clustering (complete)

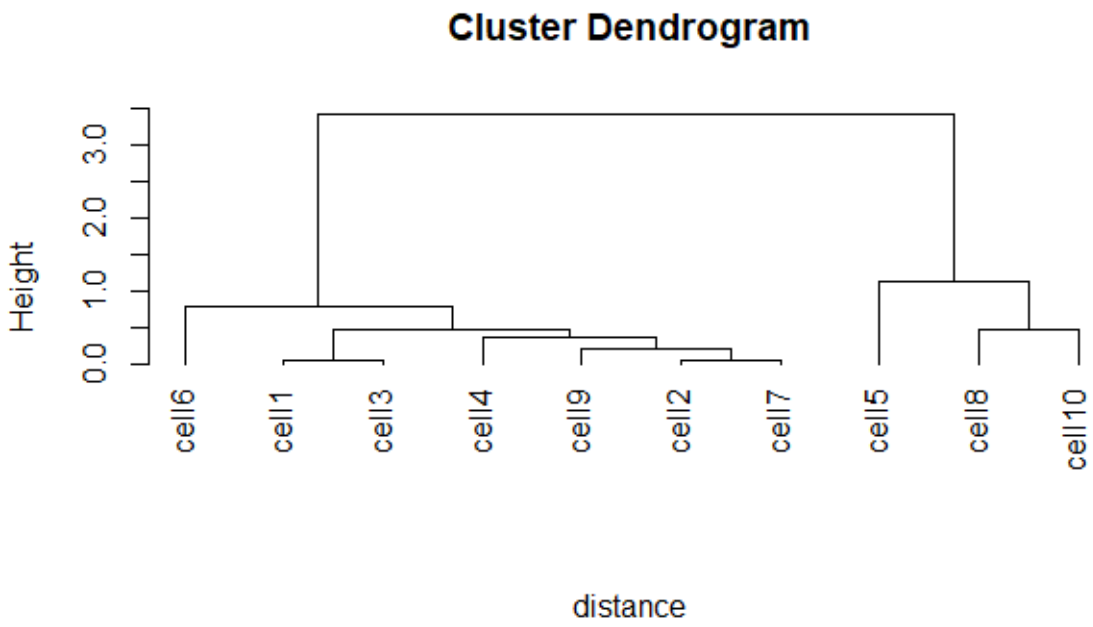


Hierarchical Clustering (complete)



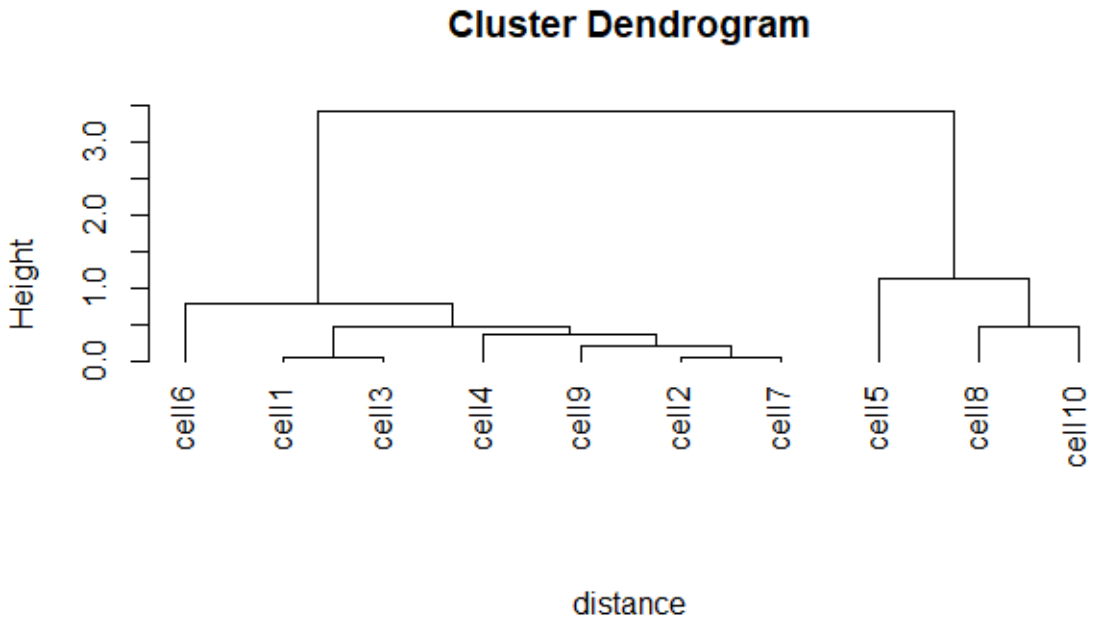
Hierarchical Clustering (average)

```
> hc.a<-hclust(distance,method="average")
> plot(hc.a,hang=-1,labels=pca$cells)
```



Hierarchical Clustering (average)

```
> hc.a<-hclust(distance,method="average")
> plot(hc.a,hang=-1,labels=pca$cells)
> member.c<-cutree(hc.c,3)
> member.a<-cutree(hc.a,3)
> table(member.c,member.a)
        member.a
member.c 1 2 3
       1 7 0 0
       2 0 1 0
       3 0 0 2
```

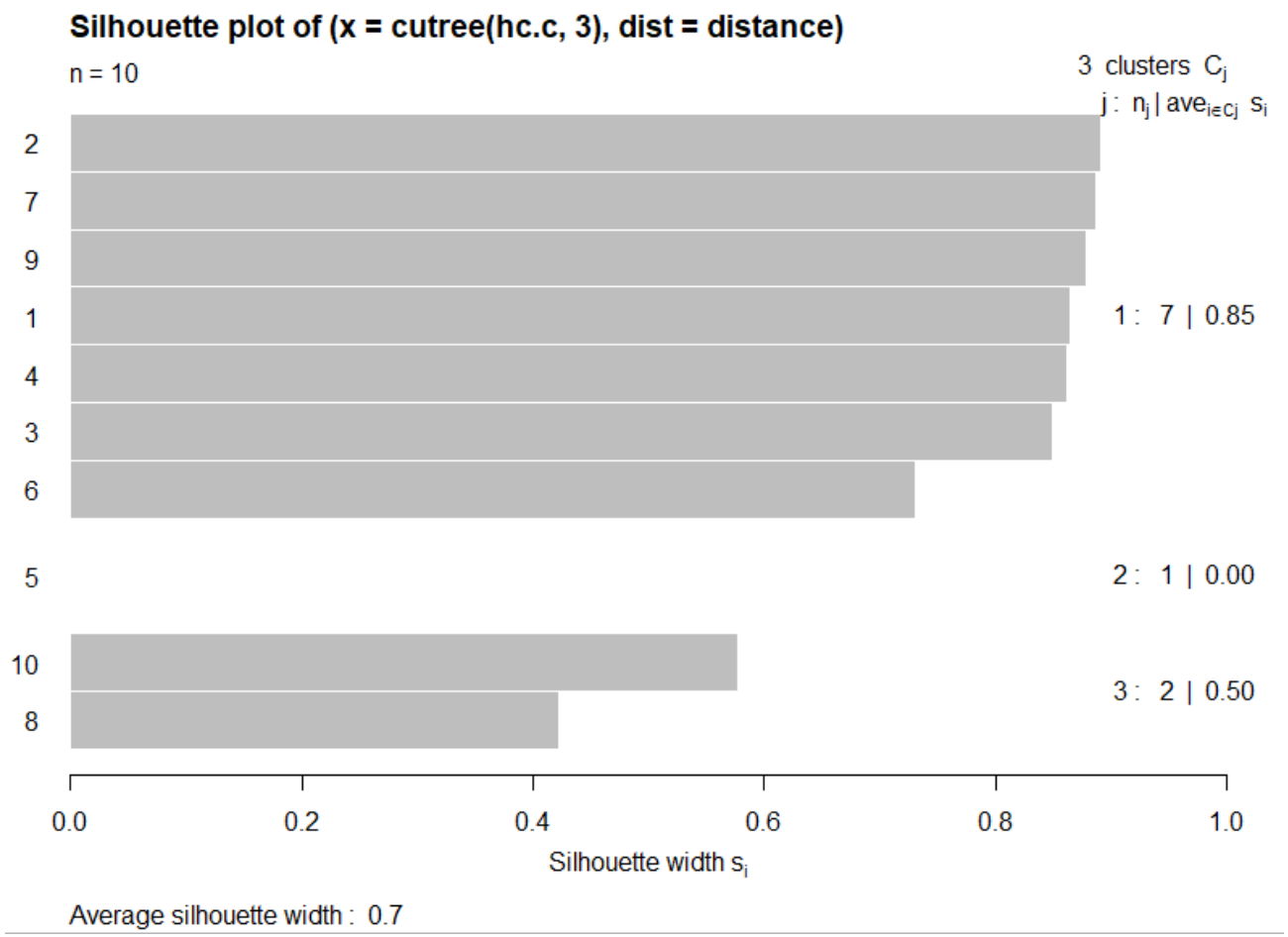


Cluster Mean

```
> aggregate(z,list(member.c),mean)
  Group.1   factor.1  factor.2
1       1  0.5827025 -0.612064
2       2 -1.9623783  1.190362
3       3 -1.0582695  1.547043

> aggregate(pca[,-c(1.1)],list(member.c),mean)
  Group.1   factor.1    factor.2
1       1  0.8443243 0.005491429
2       2 -0.3356600 0.756920000
3       3  0.0835150 0.905620000
```

Silhouette Plot

```
> library(cluster)
> plot(silhouette(cutree(hc.c,3),distance))
```



Optimal Number of Clusters

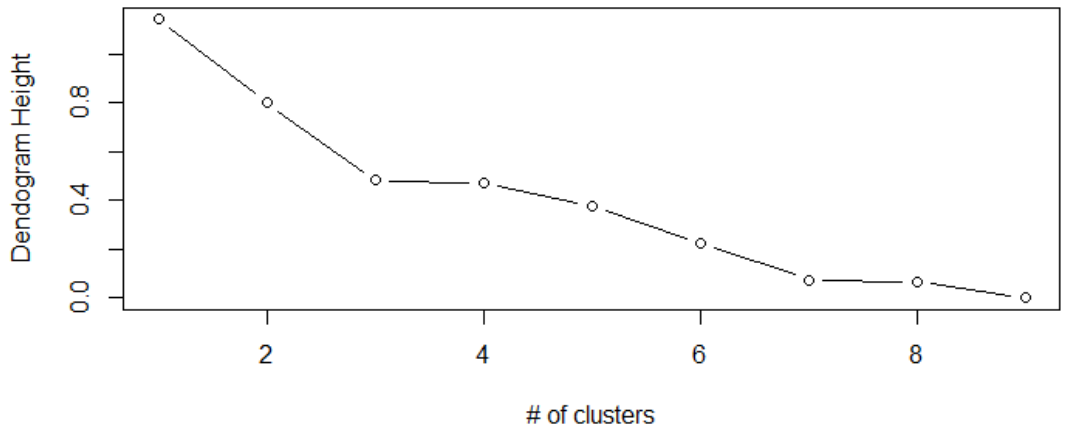
```
> Dendogram_Height=0
> for (i in 2:9) Dendogram_Height[i] <- hc.c$height[i-1]
> plot(9:1, Dendogram_Height, type="b", xlab = "# of clusters",
 ylab = "Dendogram Height")
>
```

Global Environment

Values

Dendogram_Height	num [1:9] 0 0.0655 0.0754 0.2256 0.3768 ...
distance	Class 'dist' atomic [1:45] 0.3416 0.0655 0.3555 3.1959 0.7224…
hc.a	List of 7
hc.c	List of 7

```
  merge : int [1:9, 1:2] -1 -2 -9 -4 1 -8 -6 -5 7 -3 ...
  height : num [1:9] 0.0655 0.0754 0.2256 0.3768 0.4722 ...
  order : int [1:10] 6 1 3 4 9 2 7 5 8 10
  labels : NULL
  method : chr "complete"
```



HIERARCHICAL CLUSTER IN SAS

Import Data

```
data pc;
title 'PCA Results';
input cells$ factor1 factor2;
cards;
cell1    0.93549     0.07226
cell2    0.81307     -0.0181
cell3    0.96315     0.08350
cell4    0.95191     -0.0752
cell5    -0.33566    0.75692
cell6    0.60245     0.04040
cell7    0.8073      0.01290
cell8    -0.01531    0.95305
cell9    0.8369      -0.07732
cell10   0.18234     0.85819
;
```

```
21    data pc;
22    title 'PCA Results';
23    input cells$ factor1 factor2;
24    cards;

NOTE: The data set WORK.PC has 10 observations and 3 variables.
NOTE: DATA statement used (Total process time):
      real time           0.06 seconds
      cpu time            0.04 seconds
```

Hierarchical Cluster (centroid)

```
proc cluster noeigen method=centroid rsquare nonorm out=tree data=pc;
 id cells;
 var factor1 factor2;
 run; quit;
```

PCA Results

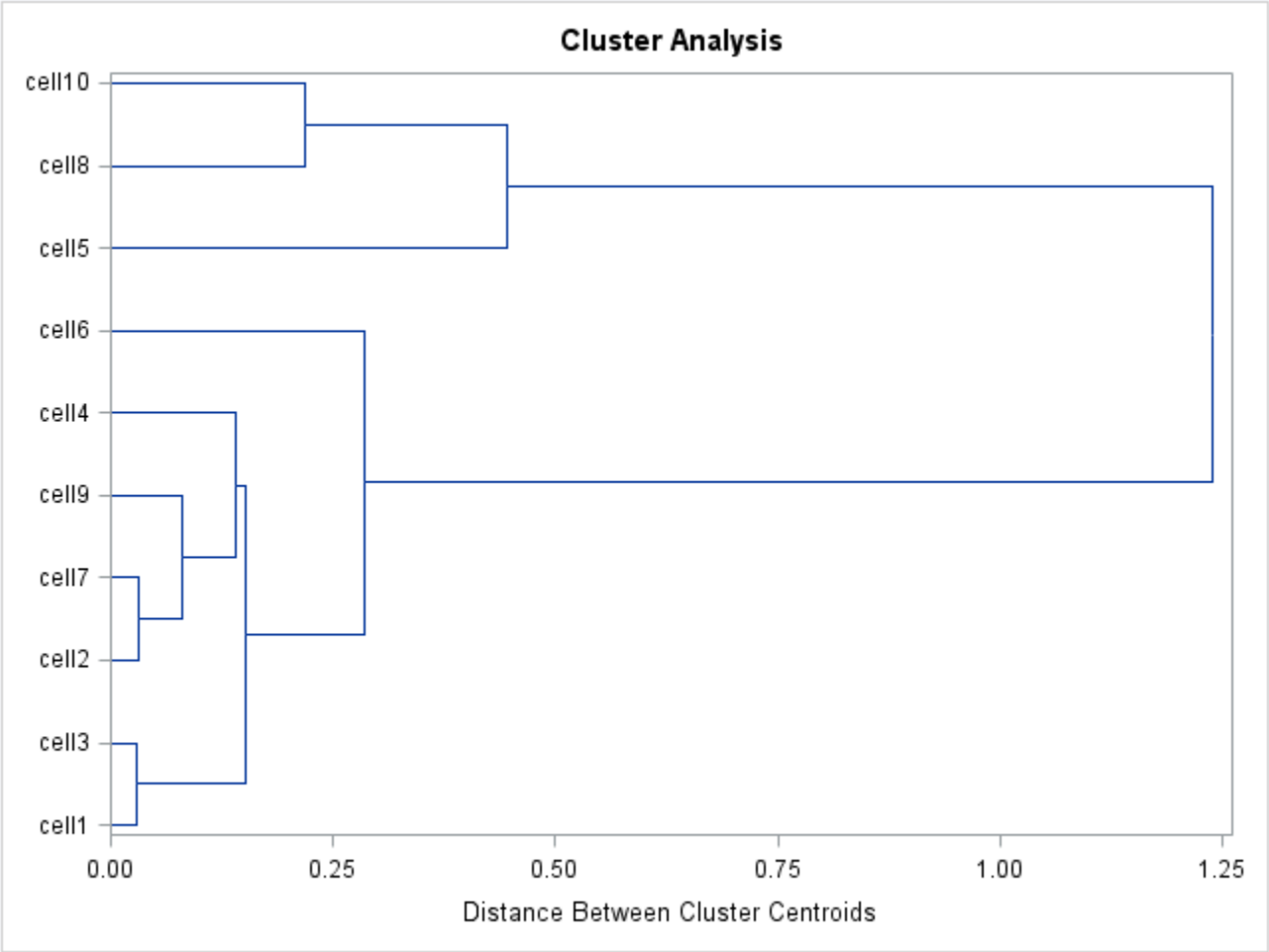
The CLUSTER Procedure
Centroid Hierarchical Cluster Analysis

Root-Mean-Square Total-Sample Standard Deviation	0.440886

Cluster History

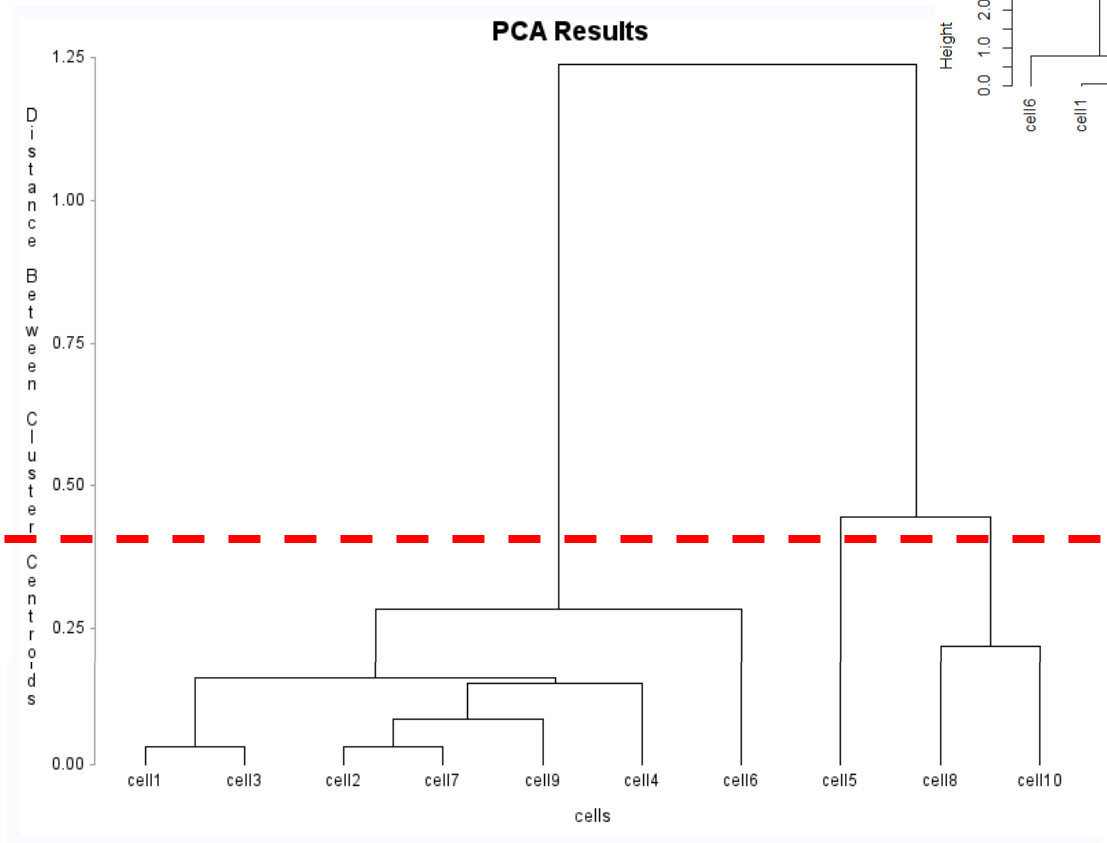
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Centroid Distance	Tie
9	cell1	cell3	2	0.0001	1.00	0.0299	
8	cell2	cell7	2	0.0001	1.00	0.0315	
7	CL8	cell9	3	0.0012	.999	0.0794	
6	CL7	cell4	4	0.0043	.994	0.1411	
5	CL9	CL6	6	0.0088	.985	0.1522	
4	cell8	cell10	2	0.0069	.979	0.2192	
3	CL5	cell6	7	0.0199	.959	0.2851	
2	cell5	CL4	3	0.0377	.921	0.4448	
1	CL3	CL2	10	0.9210	.000	1.2387	

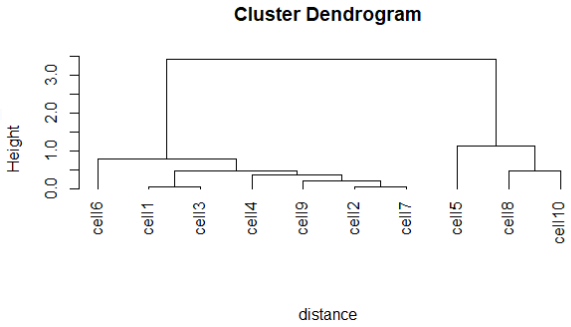
Hierarchical Cluster (centroid)



Hierarchical Cluster (centroid)

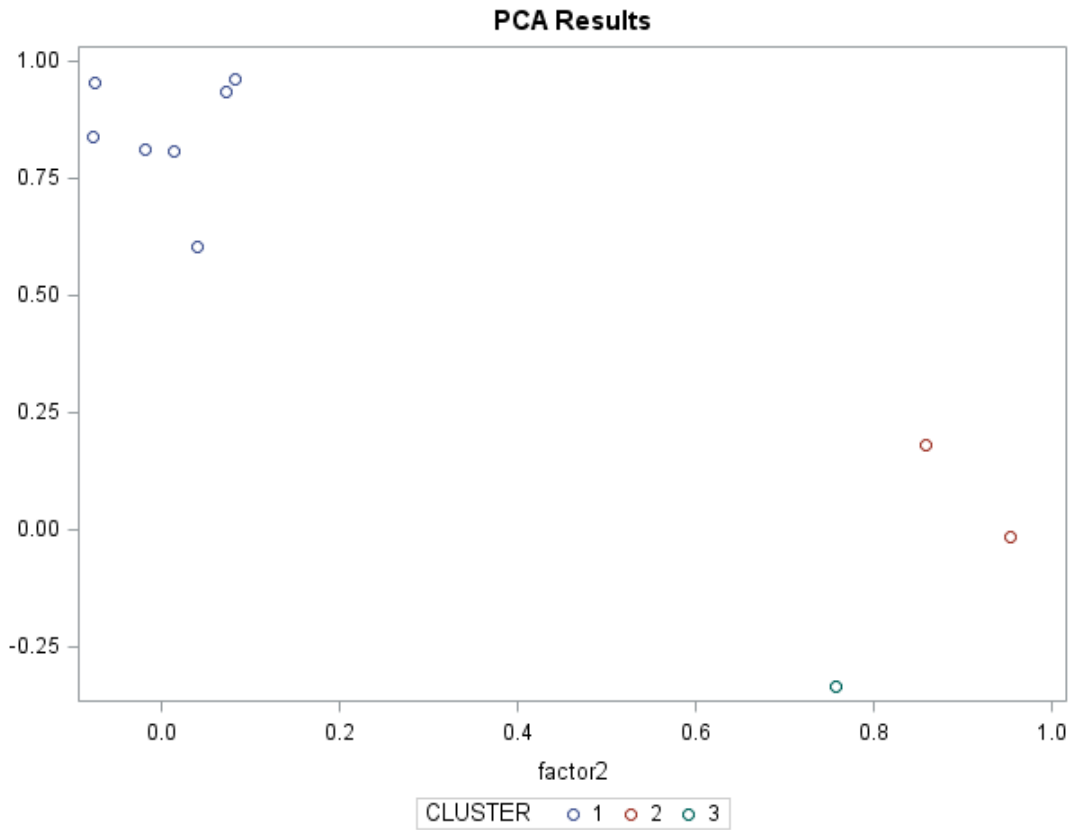
```
proc tree data=tree out=clus3 nclusters=3;
id cells;
copy factor1 factor2;
run; quit;
```





Hierarchical Cluster (centroid)

```
proc sgplot data=clus3;
scatter y=factor1 x=factor2 / group=cluster;
run; quit;
```



Hierarchical Cluster (centroid)

```
run; quit;
proc sort data=clus3; by cluster;
proc print data=clus3; by cluster;
var cells factor1 factor2;
run; quit;
```

PCA Results

CLUSTER=1

Obs	cells	factor1	factor2
1	cell1	0.93549	0.07226
2	cell3	0.96315	0.08350
3	cell2	0.81307	-0.01810
4	cell7	0.80730	0.01290
5	cell9	0.83690	-0.07732
6	cell4	0.95191	-0.07520
7	cell6	0.60245	0.04040

CLUSTER=2

Obs	cells	factor1	factor2
8	cell8	-0.01531	0.95305
9	cell10	0.18234	0.85819

CLUSTER=3

Obs	cells	factor1	factor2
10	cell5	-0.33566	0.75692

K-MEANS CLUSTER

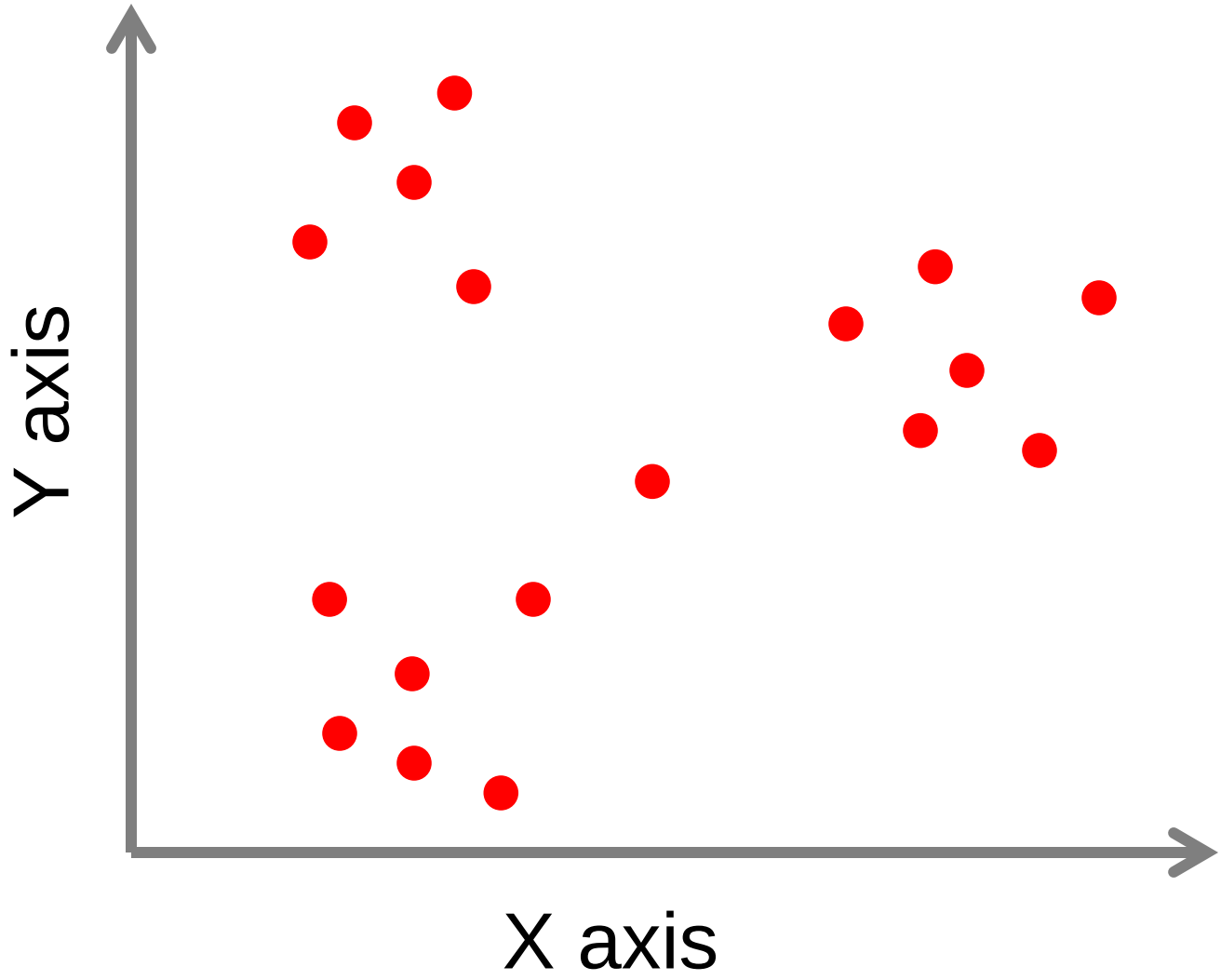
K-Means Cluster

- Most widely used for extra large data
- Observations can switch cluster membership
- Less impacted by outliers
- Multiple passes through the data allows the final solution to optimize within cluster homogeneity and between cluster heterogeneity
- Algorithm breaks the data into K clusters
- K is fixed

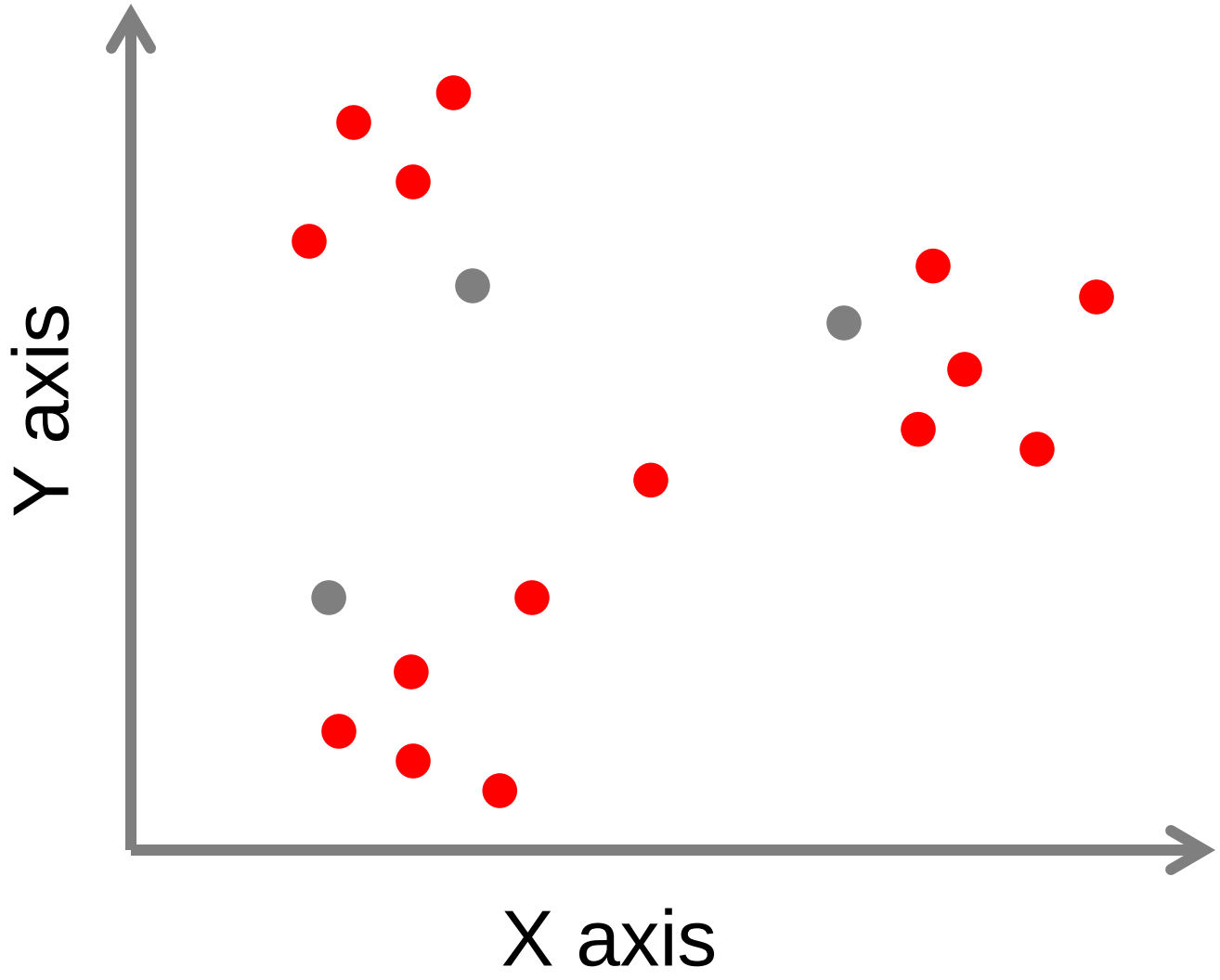
K-Means Cluster

Item	X	Y
a		
b		
c		
d		
e		
f		
g		
h		
i		
j		
Etc.		

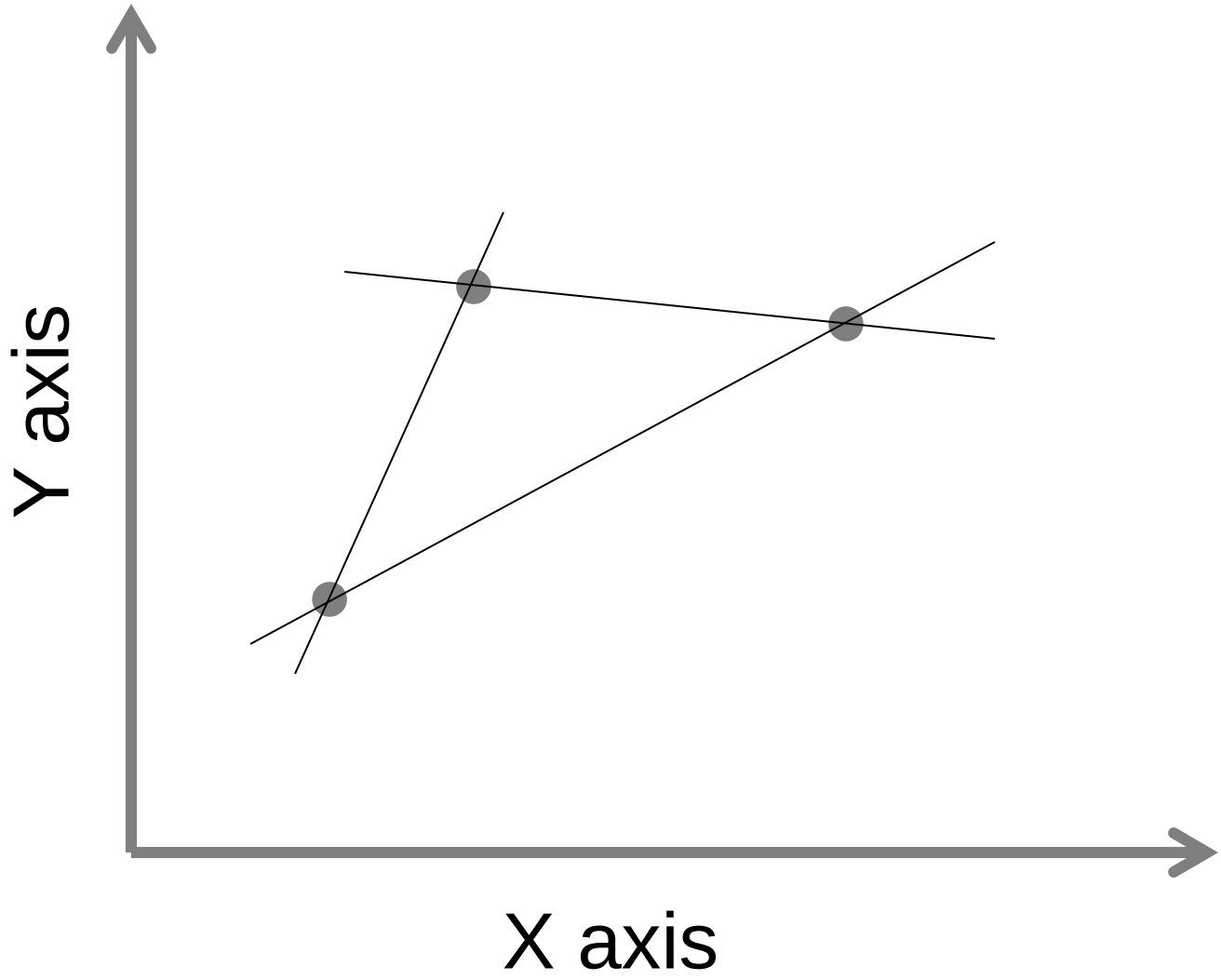
K-Means Cluster



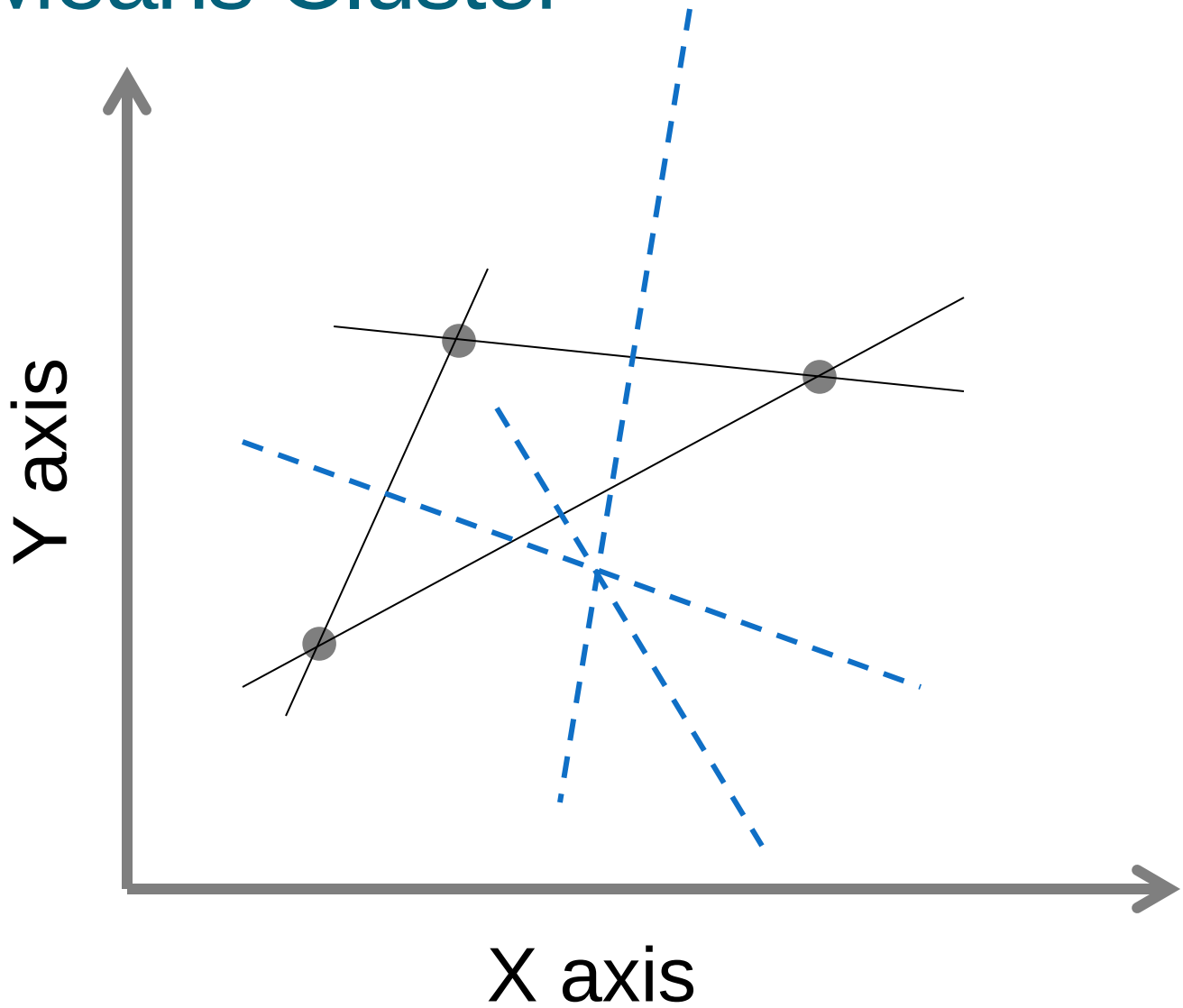
K-Means Cluster



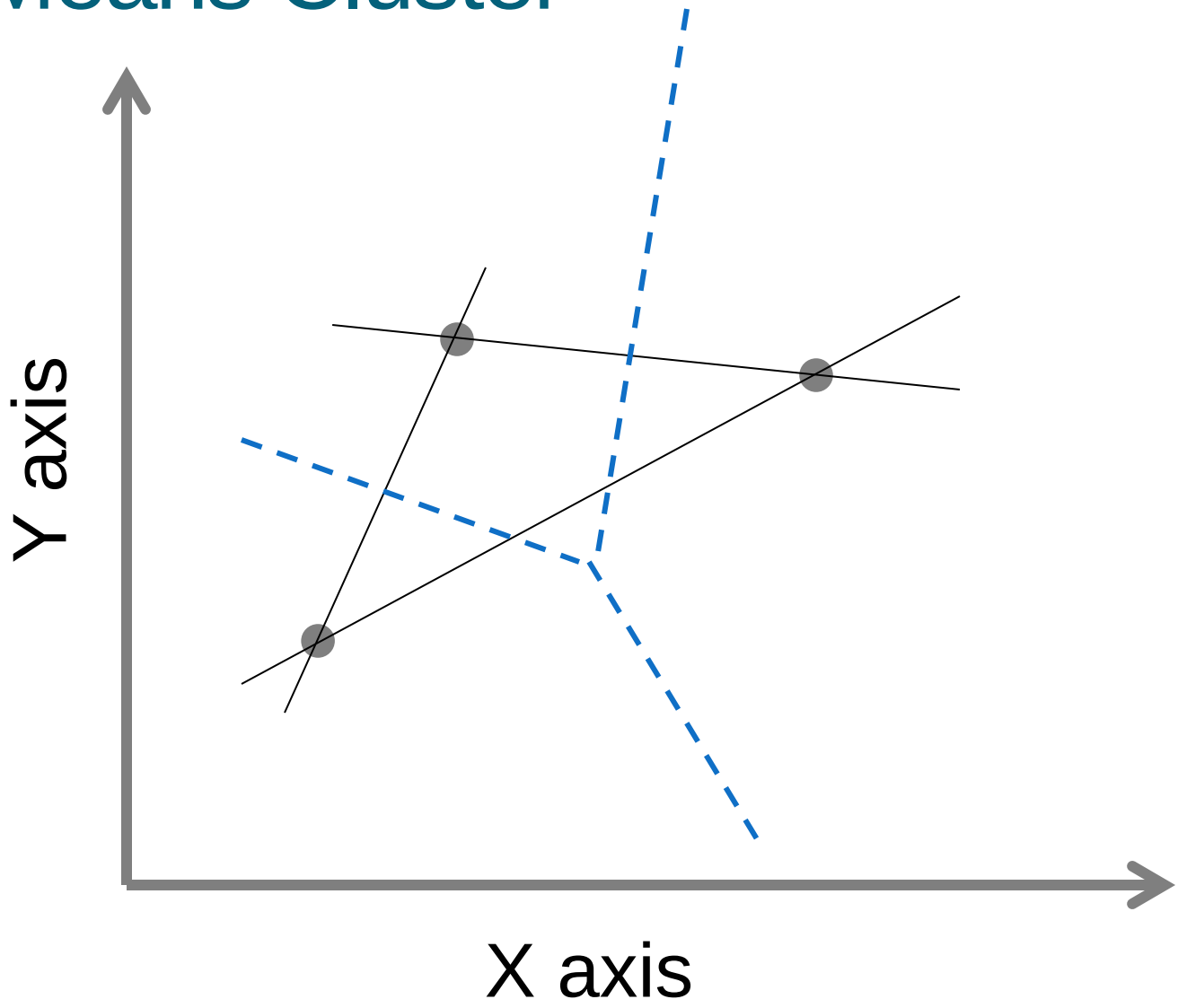
K-Means Cluster



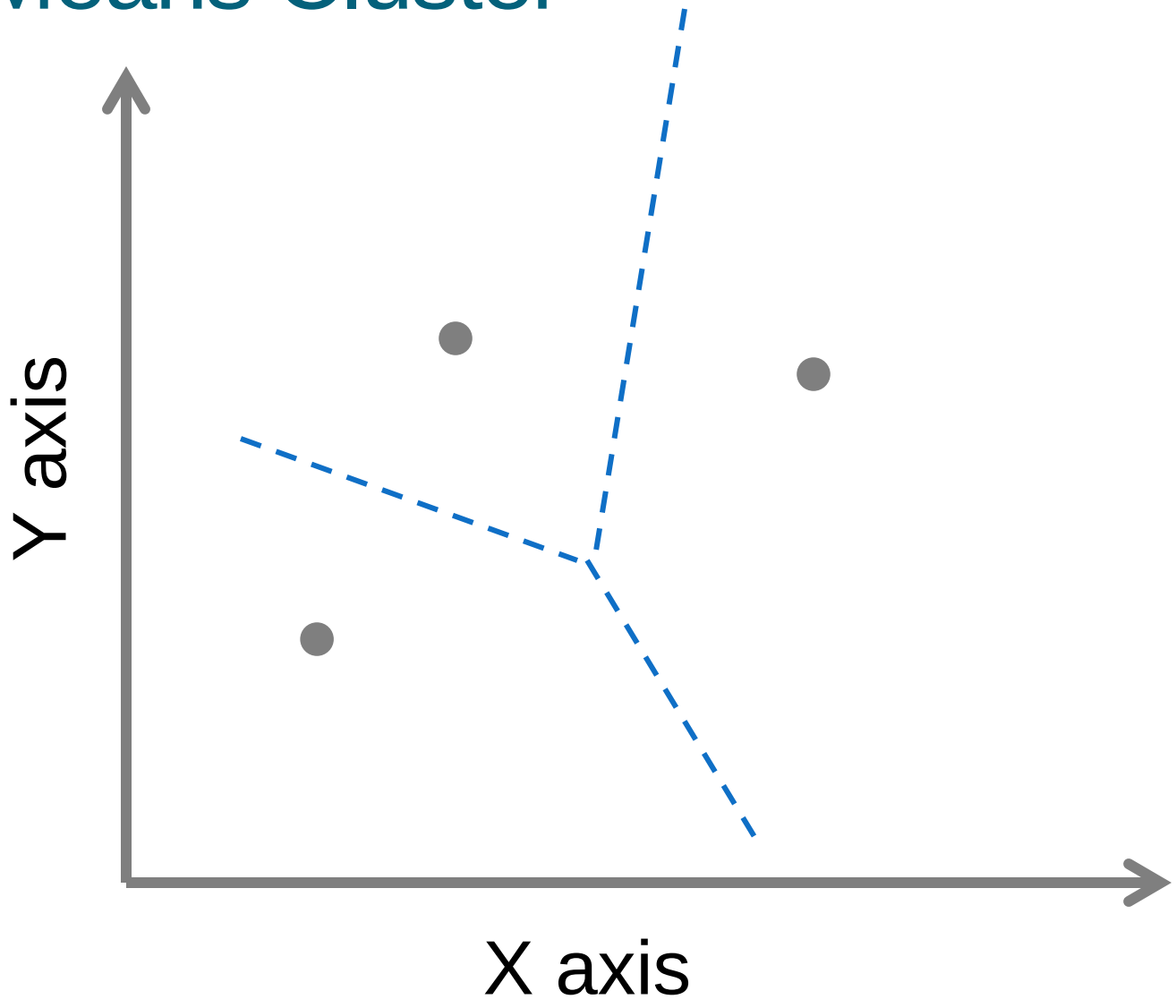
K-Means Cluster



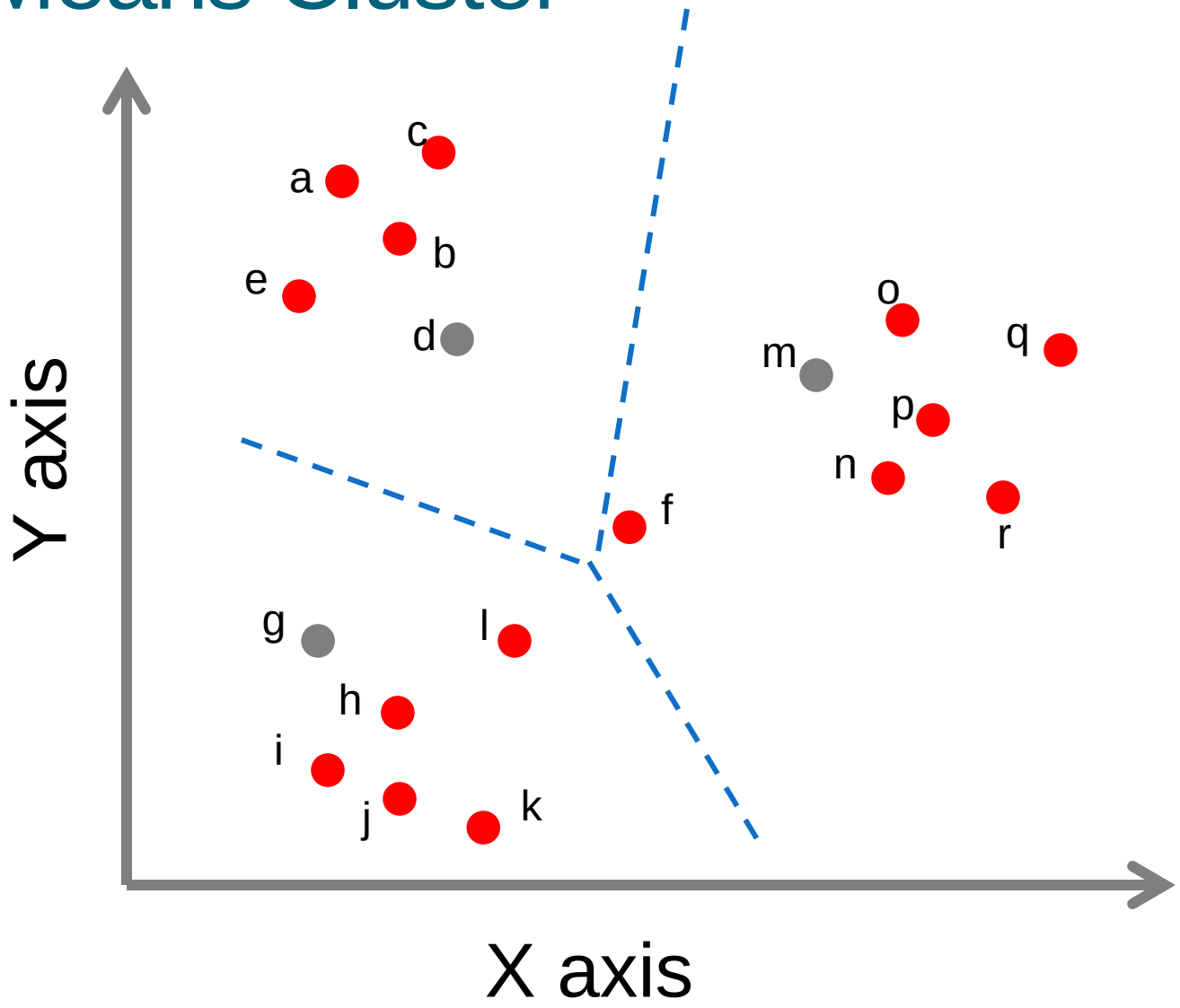
K-Means Cluster



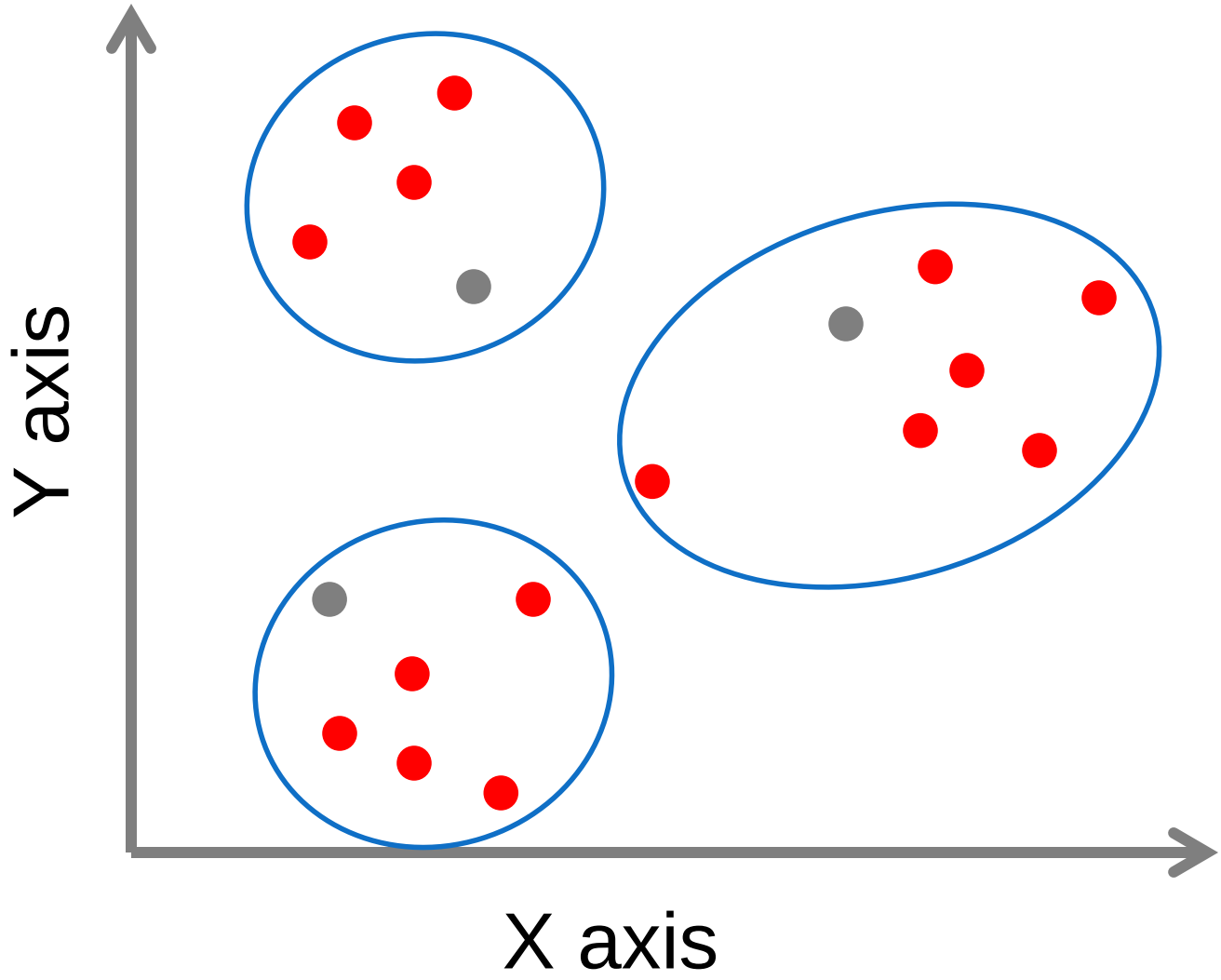
K-Means Cluster



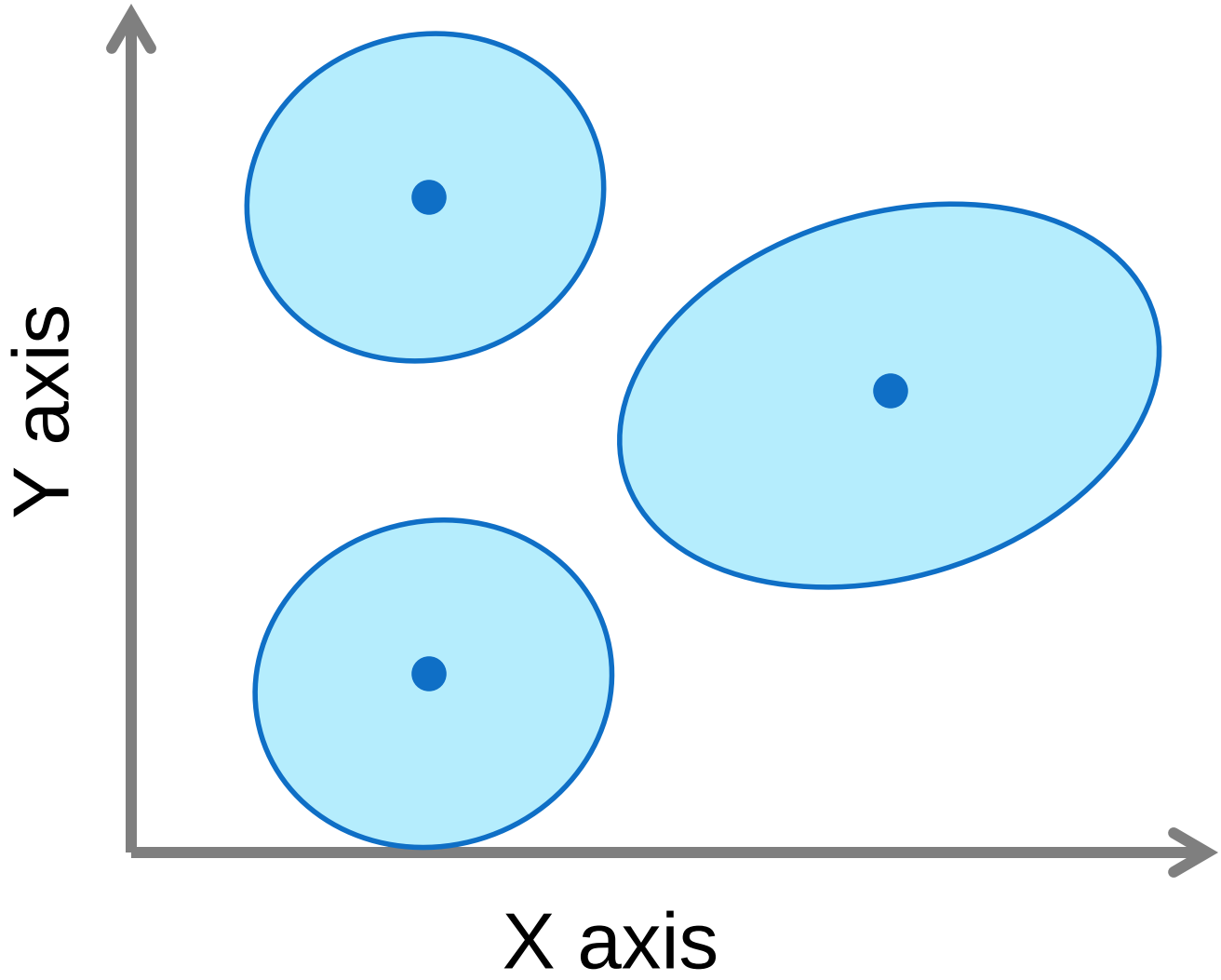
K-Means Cluster



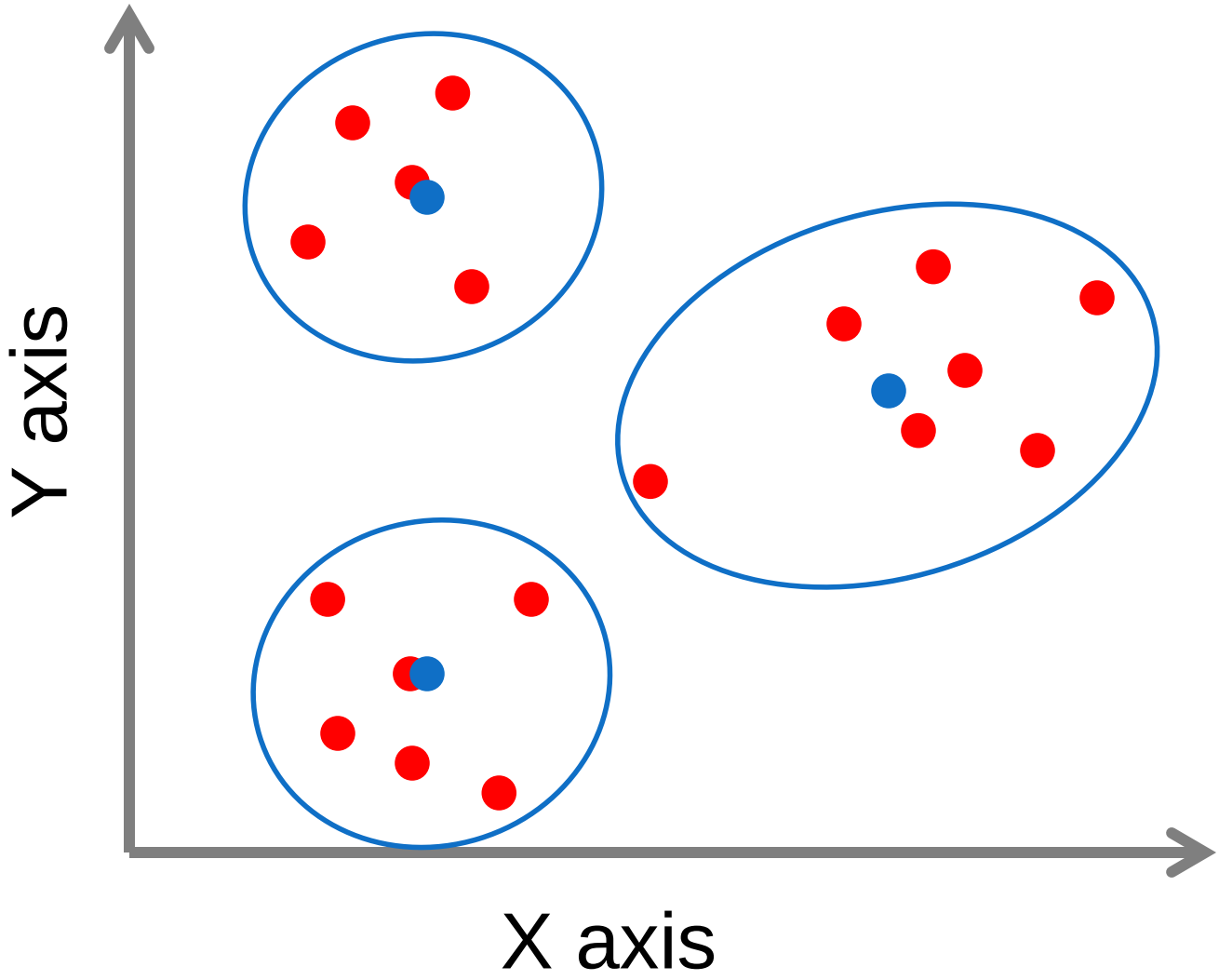
K-Means Cluster



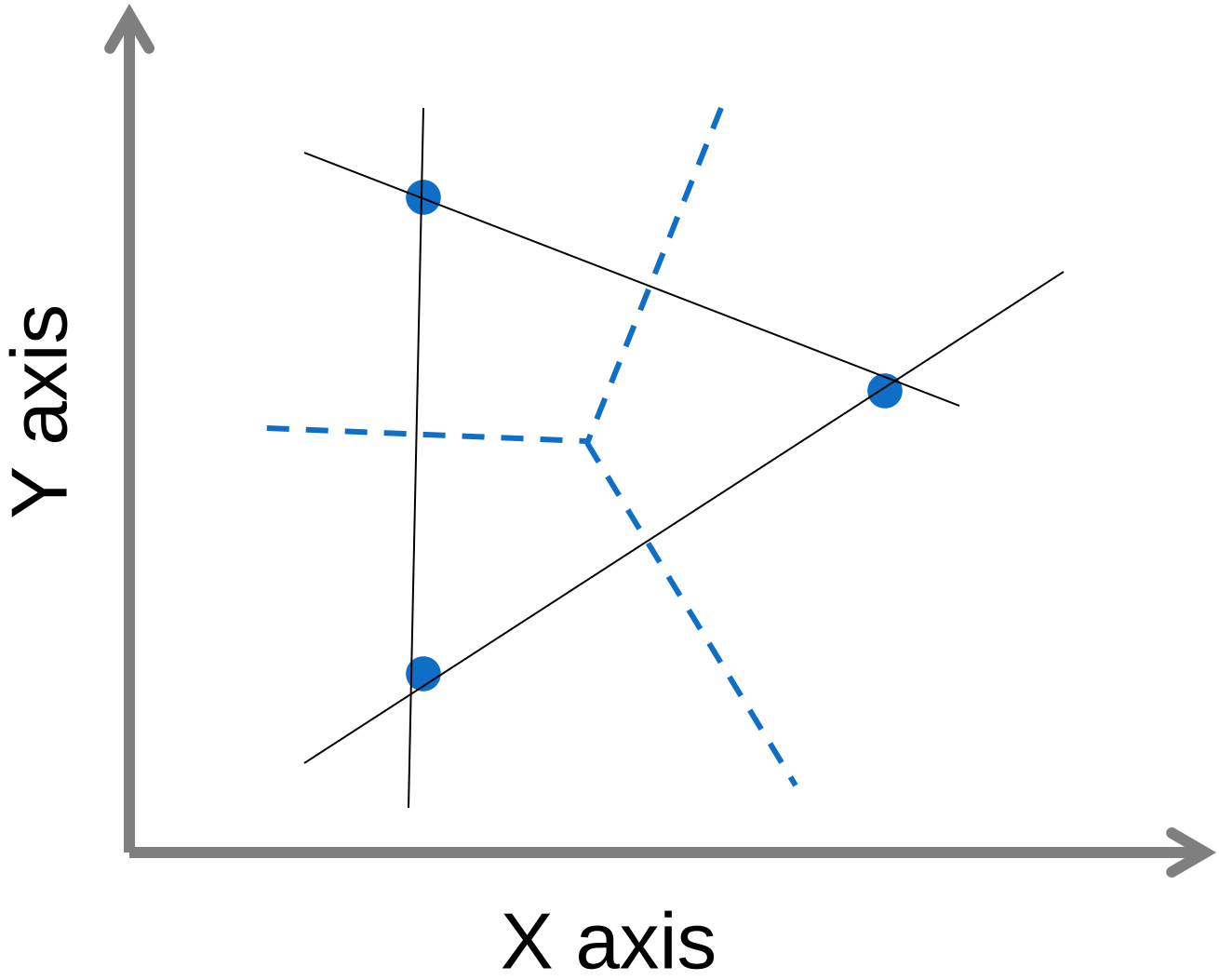
K-Means Cluster



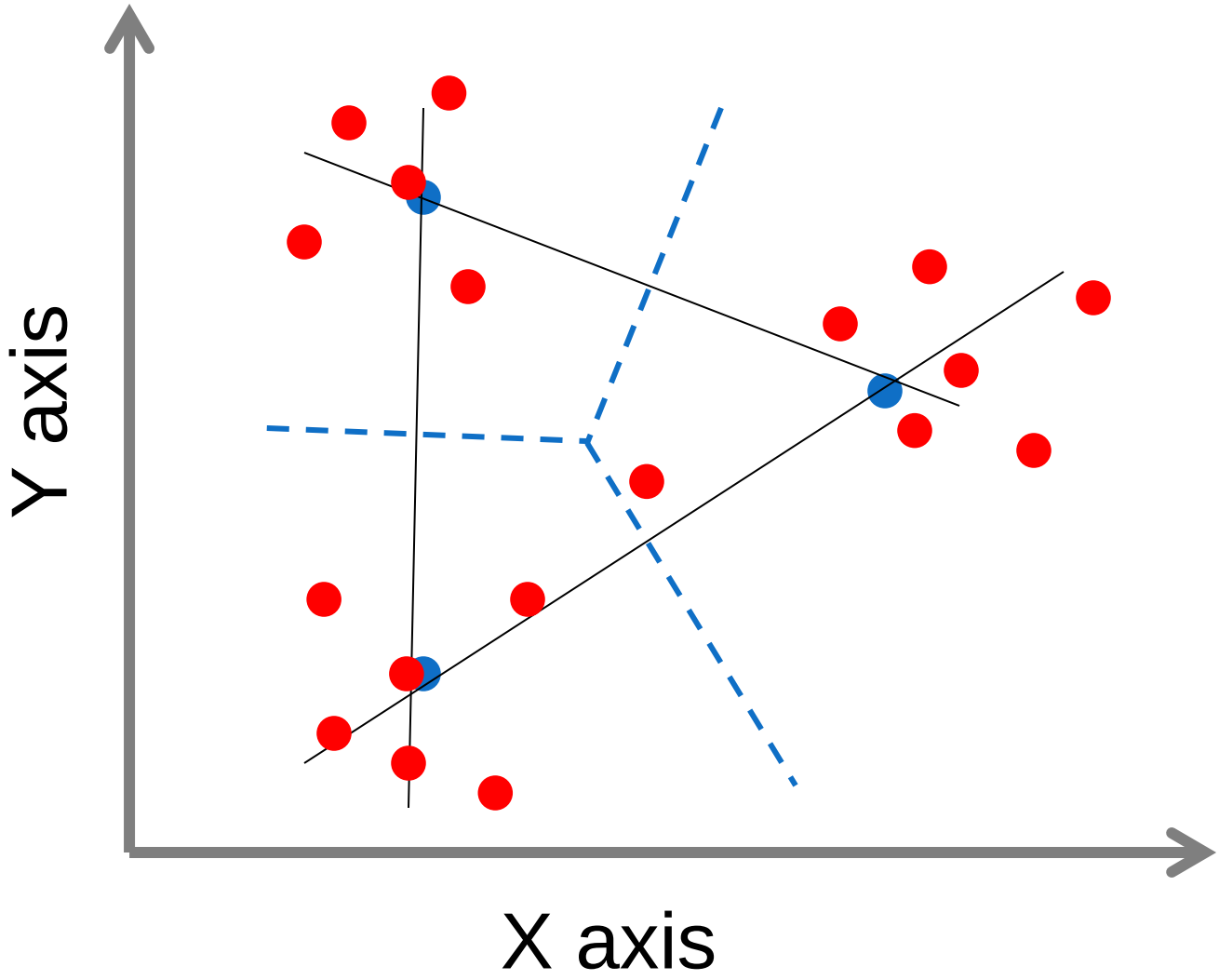
K-Means Cluster



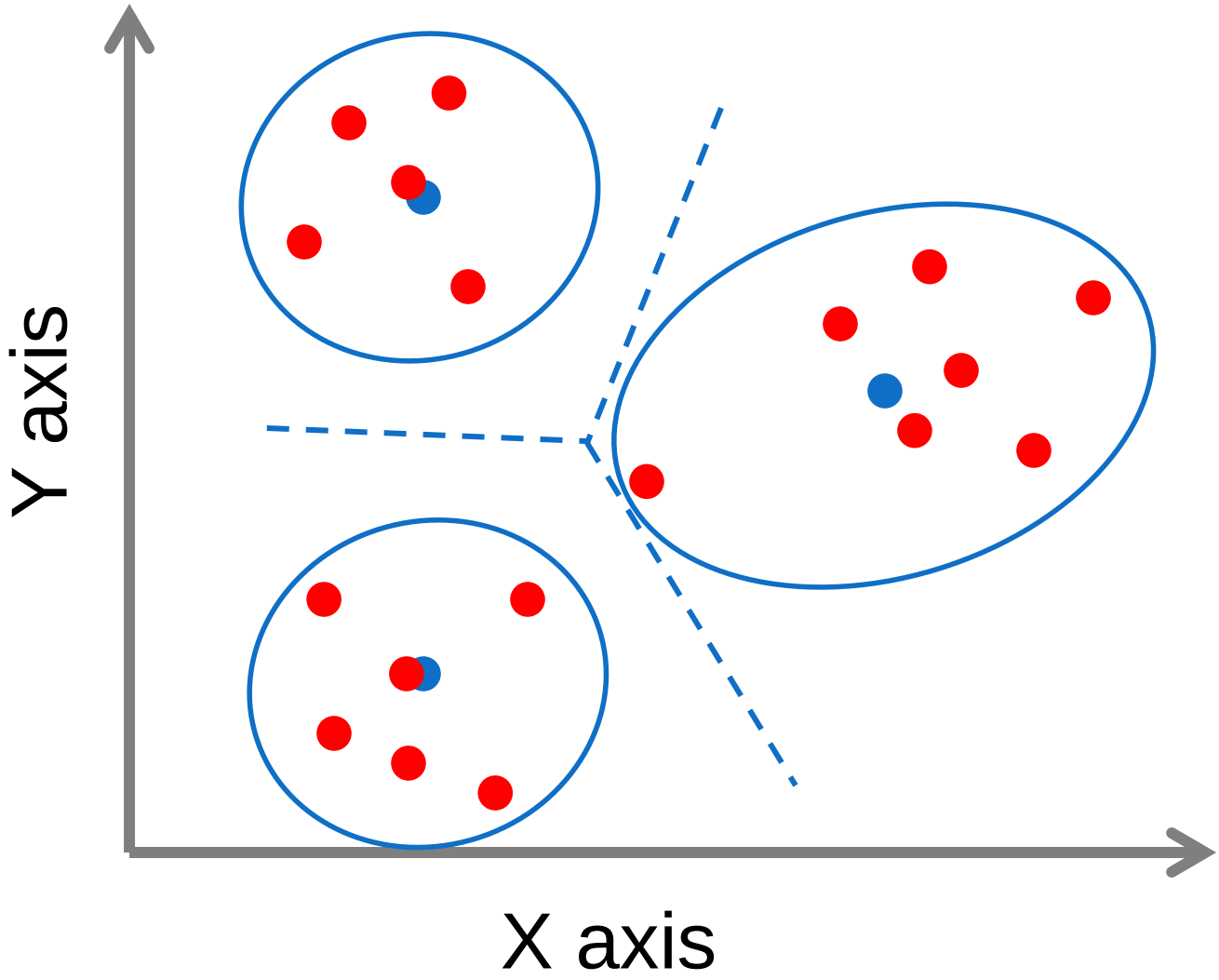
K-Means Cluster



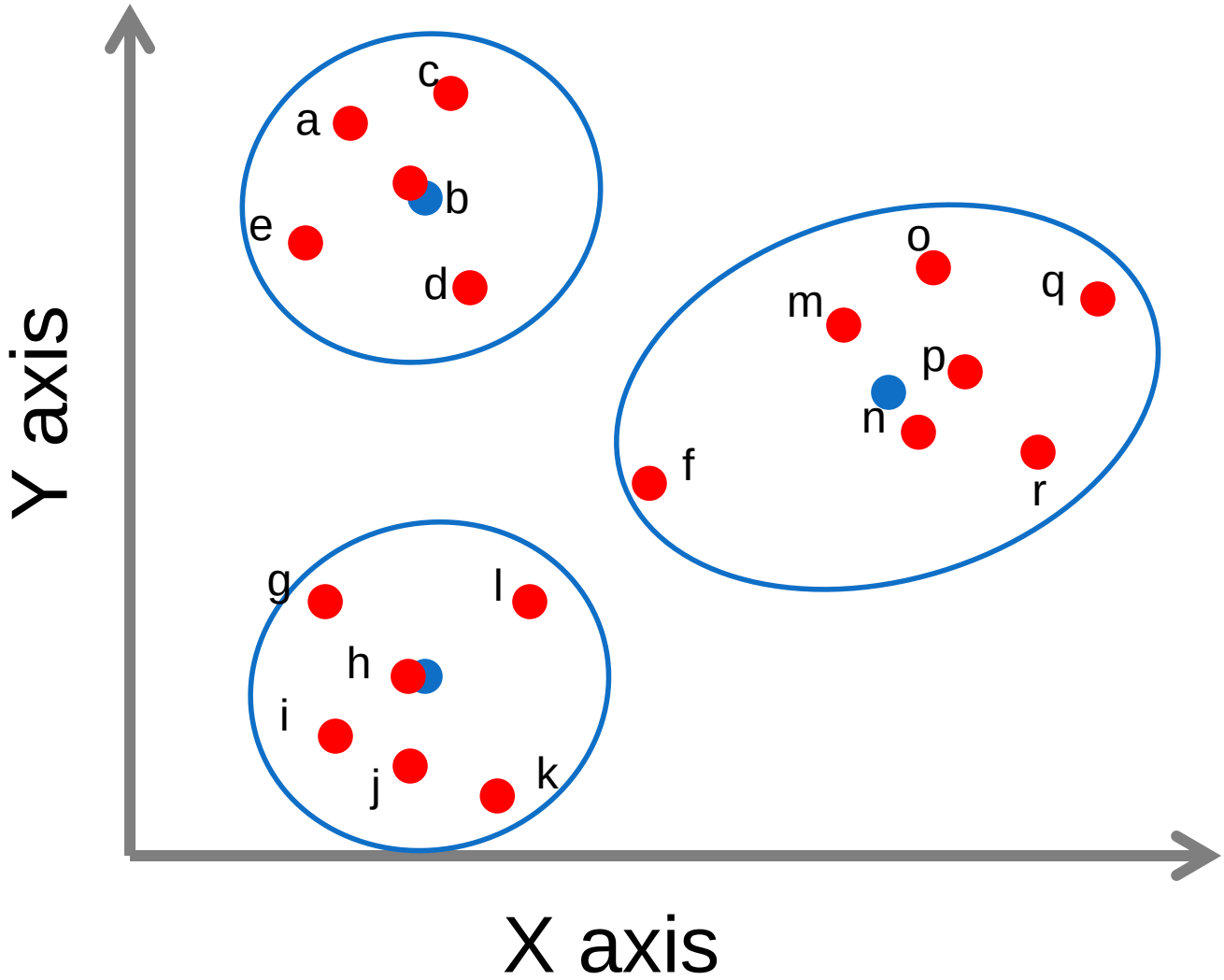
K-Means Cluster



K-Means Cluster



K-Means Cluster



Limitations of K-Means Clustering

- Underlying structure of the sample is unknown which makes it difficult to determine the number of clusters (K) needed in advance
- Poor cluster assignments cannot be modified
- Unstable solutions with a small sample (need at least 150 observations)
- Forces clusters to be round
- Outliers can distort clusters

K-MEANS CLUSTER IN R

Cell Example

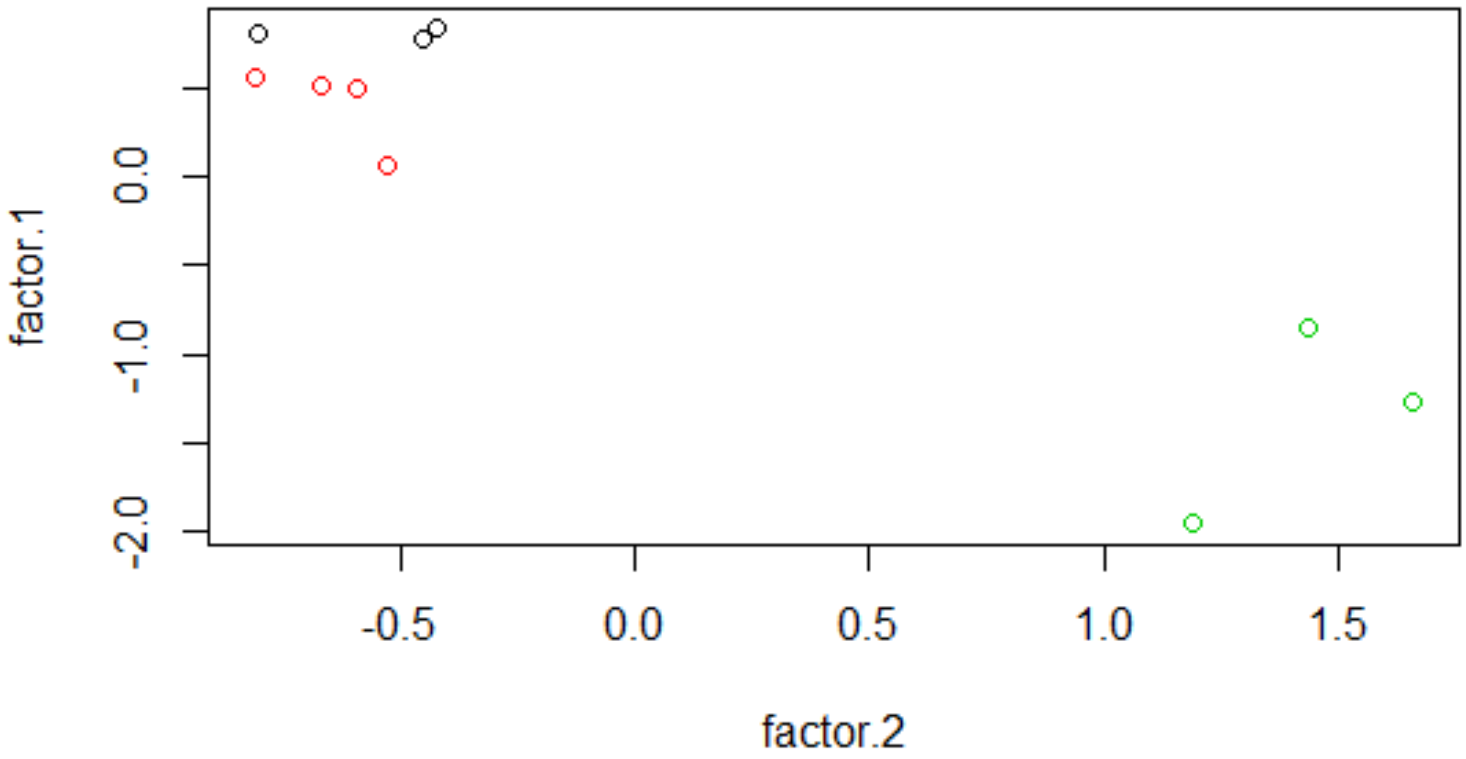
Cells	Factor 1	Factor 2
cell1	0.93549	0.07226
cell2	0.81307	-0.0181
cell3	0.96315	0.0835
cell4	0.95191	-0.0752
cell5	-0.33566	0.75692
cell6	0.60245	0.0404
cell7	0.8073	0.0129
cell8	-0.01531	0.95305
cell9	0.8369	-0.07732
cell10	0.18234	0.85819

K-Means Cluster

```
> kc<-kmeans(z,3)
> kc
K-means clustering with 3 clusters of sizes 3, 4, 3

Cluster means:
    factor.1   factor.2
1  0.8110274 -0.5608240
2  0.4114588 -0.6504941
3 -1.3596391  1.4281494

Clustering vector:
 [1] 1 2 1 1 3 2 2 3 2 3

Within cluster sum of squares by cluster:
[1] 0.0920484 0.2101224 0.7465117
 (between_SS / total_SS =  94.2 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
[8] "iter"         "ifault"
>

> kc$centers
    factor.1   factor.2
1  0.8110274 -0.5608240
2  0.4114588 -0.6504941
3 -1.3596391  1.4281494
>
```

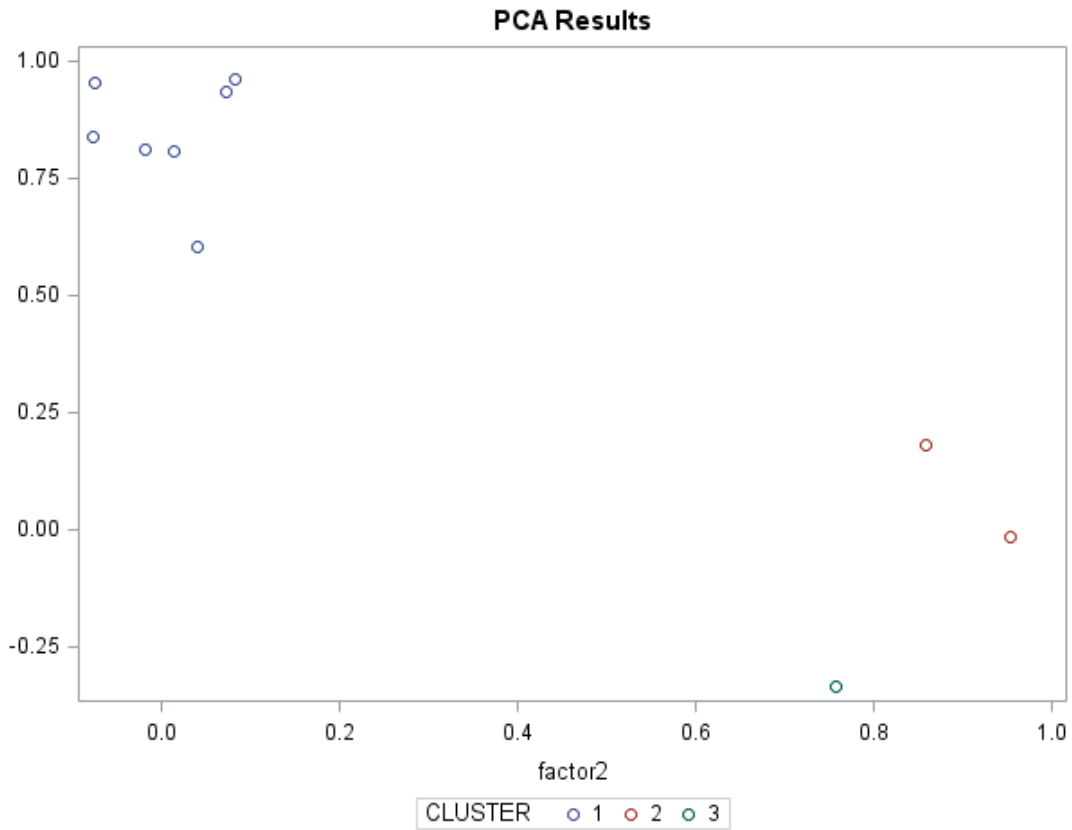
K-Means Cluster

```
> plot(factor.1~factor.2, z, col=kc$cluster)
>
```



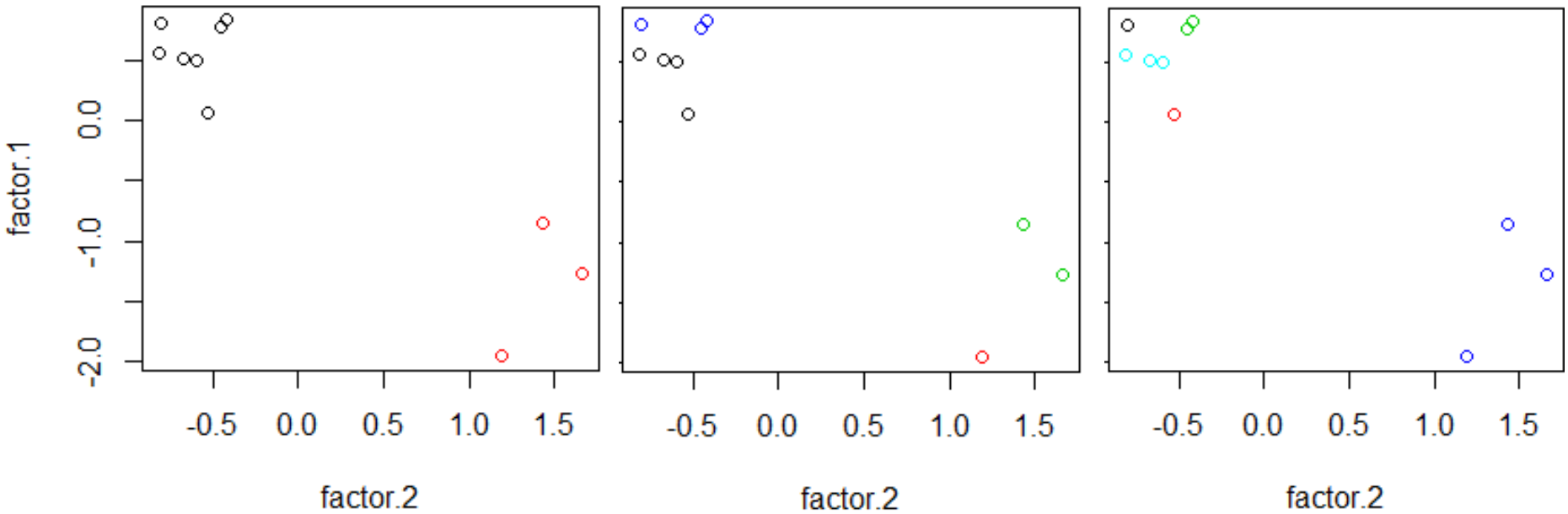
Remember?
Hierarchical Cluster (centroid)

```
proc sgplot data=clus3;
scatter y=factor1 x=factor2 / group=cluster;
run; quit;
```



K-Means Cluster

```
> plot(factor.1~factor.2, z, col=kc$cluster)
> kc<-kmeans(z,2)
> plot(factor.1~factor.2, z, col=kc$cluster)
> kc<-kmeans(z,4)
> plot(factor.1~factor.2, z, col=kc$cluster)
> kc<-kmeans(z,5)
> plot(factor.1~factor.2, z, col=kc$cluster)
>
```



K-MEANS CLUSTER IN SAS

K-Means Cluster

```
proc fastclus data=clus3 maxclusters=3 maxiter=10 list;
id cells;
var factor1 factor2;
run;
```

PCA Results

The FASTCLUS Procedure
Replace=FULL Radius=0 Maxclusters=3 Maxiter=10 Converge=0.02

Initial Seeds		
Cluster	factor1	factor2
1	-.0153100000	0.9530500000
2	0.9631500000	0.0835000000
3	-.3356600000	0.7569200000

Minimum Distance Between Initial Seeds =	0.375621

Iteration History				
		Relative Change in Cluster Seeds		
Iteration	Criterion	1	2	3
1	0.1245	0.2918	0.3784	0
2	0.0851	0	0	0

Convergence criterion is satisfied.

Cluster Listing			
Obs	cells	Cluster	Distance from Seed
1	cell1	2	0.1130
2	cell3	2	0.1421
3	cell2	2	0.0392
4	cell7	2	0.0378
5	cell9	2	0.0831
6	cell4	2	0.1345
7	cell6	2	0.2444
8	cell8	1	0.1096
9	cell10	1	0.1096
10	cell5	3	0

Criterion Based on Final Seeds =	0.0851

K-Means Cluster

Cluster Summary						
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	2	0.1096	0.1096		3	0.4448
2	7	0.1003	0.2444		1	1.1786
3	1	.	0		1	0.4448

Statistics for Variables				
Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
factor1	0.46363	0.12786	0.940844	15.904346
factor2	0.41690	0.06573	0.980667	50.724959
OVER-ALL	0.44089	0.10166	0.958648	23.182382

Pseudo F Statistic =	81.14

Approximate Expected Over-All R-Squared =	.

Cubic Clustering Criterion =	.

WARNING: The two values above are invalid for correlated variables.

GENERAL LIMITATIONS

General Limitations

- No test statistic available to validate the significance of the result
- Cluster dimensions are often randomly chosen and may not reflect real conditions → can be a statistical artifact
- Cluster analysis is powerful enough that it will provide a cluster even if no meaningful groups are embedded in the sample

General Limitations

- Choosing the variables used to group observations is the most important and different approaches may lead to different clusters
  - How to select the variable
  - Whether or not to standardize/normalize data
  - How to address multicollinearity → use PCA
    - High correlation among variables can be an issue because it may overweight other important variables
    - PCA is also controversial since low eigenvalues are dropped which may exclude factors that represent unique and important information

Best Practice

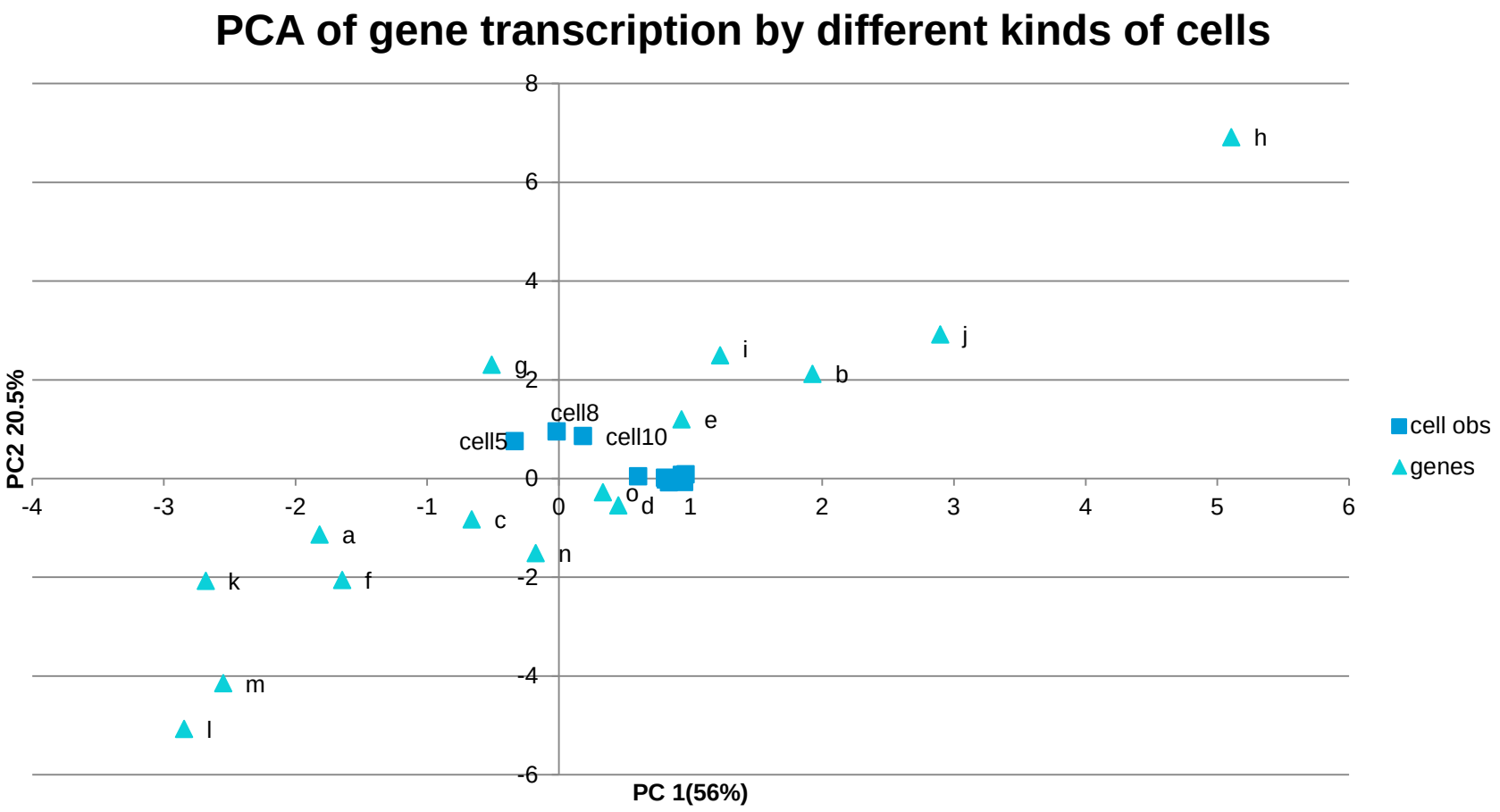
- Use Hierarchical first to determine the optimal number of clusters followed by K-Means Clustering to optimize the shape of the clusters

The End

- A good example of PCA and Cell Clustering can be seen in this paper:
- Pollen et al. (2014). Low-coverage single cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nature. Beiotech. 32:1053–1058
- doi:10.1038/nbt.2967

GENE TRANSCRIPTION

Obs	genes	cell1	cell2	cell3	cell4	cell5	cell6	cell7	cell8	cell9	cell10	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8
1	a	12	8	12	8	20	8	8	20	8	24	-1.81714	-1.13517	-0.50815	1.70426	8.9145	3.63432	-13.3285	45.435
2	b	28	28	28	28	0	8	16	12	20	16	1.92681	2.11678	-0.50975	-3.60624	-12.8978	-3.82623	17.6066	-59.070
3	c	16	16	16	12	16	16	16	16	16	12	-0.66145	-0.82913	0.36270	1.09924	4.6843	1.77916	-7.4711	19.582
4	d	20	20	20	20	8	20	20	8	24	8	0.45281	-0.54510	0.75316	-0.46999	-1.2925	0.28286	2.8269	-5.344
5	e	28	24	24	24	4	12	8	20	12	8	0.93275	1.19706	-0.77193	-1.76390	-3.8704	-2.37981	11.3183	-33.983
6	f	4	32	12	12	28	0	8	16	16	4	-1.64457	-2.06041	-0.54662	5.87948	8.5946	3.09480	-13.5101	45.141
7	g	18	12	18	12	30	12	12	30	12	36	-0.50823	2.03144	-0.08978	-0.49679	0.6718	1.35217	-1.0036	5.940
8	h	42	42	42	42	0	12	24	18	30	24	5.10768	6.90936	-0.09217	-8.46254	-32.0466	-9.83865	45.3991	-150.816
9	i	24	24	24	18	24	24	24	24	24	18	1.22530	2.49050	1.21650	-1.40432	-5.6734	-1.43056	7.7824	-32.839
10	j	30	30	30	30	12	30	30	12	36	12	2.89668	2.91655	1.80218	-3.75816	-14.6387	-3.67502	23.2295	-70.228
11	k	8	12	8	8	20	0	4	28	4	20	-2.67918	-2.07466	-1.31858	4.19442	15.2148	2.16522	-22.1464	74.130
12	l	7	7	7	7	0	2	4	3	5	4	-2.84450	-5.07210	-1.13611	3.67821	15.8254	5.19239	-24.0821	78.550
13	m	8	8	8	6	8	8	8	8	8	6	-2.54819	-4.14876	-0.49110	3.60280	15.0420	4.98888	-22.7247	72.003
14	n	15	15	15	15	12	30	30	12	36	12	-0.17512	-1.51687	2.04818	0.98160	4.7955	-0.51952	-7.6067	24.694
15	o	21	21	21	21	0	6	12	9	15	12	0.33637	-0.27951	-0.71854	-1.17809	-3.3234	-0.82002	3.7104	-13.196


PCA of gene transcription by different kinds of cells
PC2 20.5%
PC 1(56%)
8
6
4
2
0
-2
-4
-6
-4
-3
-2
-1
0
1
2
3
4
5
6
h
j
i
b
g
e
cell8
cell5
cell10
o
d
c
a
n
k
f
m
l
cell obs
genes

Monotheic

Polythetic

Divisive Clustering