Module 2: Introduction to Statistics

Niko Kaciroti, Ph.D.
BIOINF 525 Module 2: W15
University of Michigan

Topic

- Multiple Testing
  - Family-Wise Error Rate
    - Bonferroni adjustment
  - False Discovery Rate
- Multivariate Data Analysis
  - Principle Component
    - Screen plot, Bipolar plot
  - Cluster Analysis: K-means, Hierarchical Clustering
    - Dendrograms
    - Heatmaps

Multiple Testing

- Multiple testing is frequently used in large data sets, particularly in discovery science
  - E.g. Genome-wide association studies (GWAS) – test up to several million genetic variants for association with a trait
- When many tests are performed the following questions are relevant:
  - Are there any true positive results?
  - How many are false positive?
  - Which are the true positives?

Type I Error: False Positive (FP) and True Positive (TP) Rate

- Type I error or False Positive Rate for testing a single hypothesis usually is set at α = 0.05

$$FP = Pr(\text{Reject } H_0 \mid H_0 \text{ is true}) = .05$$
$$TN = Pr(\text{Do not Reject } H_0 \mid H_0 \text{ is true}) = .95\ (= 1 - \text{FP})$$

- Suppose we are testing two independent null hypotheses H_{01} and H_{02}. In such case the Type I Error is compounded

$$Pr(\text{Reject } H_{01} \text{ or } H_{02} \mid H_{01} \text{ and } H_{02} \text{ are true}) = 1 - (.95)^2 = .0975$$

- In general, the more hypotheses you test, the more likely it is to see by chance a difference that is not there

Family-Wise Error Rate (FWER)

- Suppose we test a family of hypotheses, *e.g.* we collected 10 possible biomarkers for high BP, and test whether:
  - Marker 1 does (or does not H_{01}) relate with high BP
  - Marker 2 does (or does not H_{02}) relate with high BP
  - Marker *m* · · · · · ·
- If we perform all the tests, and they are independent, the probability that we make at least one false positive (or "false discovery") is around 0.4 $(= 1 - (.95)^{10})$

Family-Wise Error Rate (FWER)

- Similarly, if we test if a treatment effects 10 outcomes (BP, diabetes,…,lung cancer), the probability of making at least one false positive is still around 0.4
- This is called the "family-wise error rate" (FWER):
  - FWER=*Pr(Reject at least one of* H_{0k} *|All* H_{0k} *are true)*
- FWER is always greater than α = 0.05 and could be quite large if number of tests, *m*, is large

Controlling Family-Wise Error Rate (FWER): Bonferroni Adjustment

- Let H_{0k} be a family of *k=1,2,…,m* hypotheses. If we reject H_{0k} when $p_k < \alpha$, then the following is true:

$$\text{FWER}=Pr(\text{Reject at least one of } H_{0k} \mid \text{All } H_{0k} \text{ are true}) \leq m\alpha \qquad (1)$$

- From Eq. (1), if we carry out the significance of each test at $p_k < \alpha^*$ where $\alpha^* = \frac{\alpha}{m}$, then the FWER is at most:

$$\text{FWER} \leq m\alpha^* = m\frac{\alpha}{m} = \alpha$$

Controlling Family-Wise Error Rate (FWER): Bonferroni Adjustment

- Bonferroni adjustment: If α = 0.05 and there are m=10 tests, then use $\frac{\alpha}{10}$ =.005 as a criteria to reject a null hypothesis, i.e. p < .005
- Bonferroni adjustment works OK for classical multiple testing (when m ~ 3-5). But in general it is too conservative. It overprotects against FWER and, as a result, the Power is reduced.
- For a large number of multiple testing, the False Discovery Rate (FDR) method is a better alternative

False Discovery Rate (FDR)

- **FDR** is the expected rate of false discoveries among all discoveries (rejected null hypotheses)

$$FDR = \frac{\#False\ Discoveries}{\#All\ Discoveries}$$

  - E.g. If there were m=1000 discoveries (1000 null hypotheses were rejected) and a FDR level (q-value) for these tests was 0.05, then 50 among 1000 discoveries were expected to be false discovery

- How to adjust for multiple testing so that FDR ≤ .05?
  - For each of the *m* tests, get the p-value. Order them: $p_1 \leq p_2 \leq \ldots \leq p_m$. Find the largest k, such that $p_k \leq \frac{k*.05}{m}$, then reject $H_{01},\ldots,H_{0k}$

Illustration from the Example in Benjamini et. al. Article on FDR

Benjamini, Yoav; Hochberg, Yosef (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, *Series B (Methodological)* 57 (1): 289–300.

http://www.math.tau.ac.il/~ybenja/MyPapers/benjamini_hochberg1995.pdf

Example: 15 tests resulted in the following 15 p-values:

0.0001, 0.0004, 0.0019, 0.0095, 0.0201, 0.0278, 0.0298, 0.0344, 0.0459, 0.3240, 0.4262, 0.5719, 0.6528, 0.7590, 1.000.

Example: Bonferroni and FDR Adjustment (m=15 tests)

k	p_k(order)	$\alpha = .05$
1	.0001	<.05
2	.0004	<.05
3	.0019	<.05
4	.0095	<.05
5	.0201	<.05
6	.0278	<.05
7	.0298	<.05
8	.0344	<.05
9	.0459	<.05
10	.3240	>.05
	….	…
15	1.000	>.05

Example: Bonferroni and FDR Adjustment (m=15 tests)

k	p_k(order)	$\alpha = .05$	Bonferroni $\alpha = \frac{.05}{m} = .0033$	FDR(q=.05) $\frac{k*.05}{m}$
1	.0001	<.05	<.0033	<.0033
2	.0004	<.05	<.0033	<.0066
3	.0019	<.05	<.0033	<.0099
4	.0095	<.05	>.0033	<.0132
5	.0201	<.05	>.0033	>.0165
6	.0278	<.05	…	>.0198
7	.0298	<.05		>.0231
8	.0344	<.05		>.0264
9	.0459	<.05		>.0297
10	.3240	>.05		>.0330
	….	…	…	…
15	1.000	>.05	>.05	>.05

- FDR: k=4 is the largest k for which $p_k \leq \frac{k*.05}{m}$. Thus, reject for p_1, p_2, p_3, p_4

Topic

- Multiple Testing
  - Family-Wise Error Rate
    - Bonferroni adjustment
  - False Discovery Rate

- **Multivariate Data Analysis**
  - **Principle Component**
    - Screen plot, Bipolar plot
  - Cluster Analysis: K-means, Hierarchical Clustering
    - Dendrograms
    - Heatmaps

Multivariate Analysis

- Multivariate analysis is different from the other modeling techniques (e.g. t-test, regressions) because there is no outcome or predictor

- In multivariate statistics we look for **structure in the data**

- Two common methods that look for structure are:
  - *Principal Component Analysis:* Look for structure among variables
  - *Cluster Analysis:* Look for structure among individuals

Multivariate Analysis

- In multivariate data the number of variables of interest ($X_1, X_2, \ldots, X_p$) may be large or too large (e.g. high dimensional data)
  - This may cause problems with statistical modeling (*i.e.* regression)
    If p > n then the degrees of freedom (*df=n-p-1*) for regression would be negative.
    In that case one can't run a multiple regression (Not enough data points (n) to estimate *p* parameters)
  - The interpretation of large data or results will be cumbersome
  - There may be multiple testing issues (e.g. many false discoveries)
- Thus, data reduction when dealing with multivariate data is needed

Principal Components Analysis (PCA)

- ***Principal component analysis*** (PCA) is a dimension-reduction method that generates a new set of decorrelated variables
- The new variables, called Principal Components (PC), are linear combinations of the original variables ($X_1, X_2, \ldots, X_p$)
- The idea of PCA is to find a small number of **linear combinations** of the variables (X's), which capture most of the variation of the original data

Principal Components Analysis (PCA)

- Simple example: Suppose, that you had four measures (i.e. exam scores in math, biology, physics, chemistry). How would you summarize overall performance into a single score?
- A solution is to take the mean of the four variables

$$S = \frac{x_1+x_2+x_3+x_4}{4} = \frac{1}{4}x_1 + \frac{1}{4}x_2+\frac{1}{4}x_3+\frac{1}{4}x_4$$

- S is a linear combination of x_1, x_2, x_3, x_4 with coefficient *l* = (1/4, 1/4, 1/4, 1/4)
- PCA is statistical technique that finds few linear combinations (similar to S) that summarize the data

Principal Component Analysis (PCA)

Original Data: $X_1, \quad X_2, \quad X_3, \quad X_4, \quad \dots \quad , X_p$

PCA

PC: $PC_1, \quad PC_2, \quad \dots \quad , PC_k$ $\quad k \leq p$

- $PC = l_1X_1 + l_2X_2 + \cdots + l_pX_p$
- $l_1, l_2, \dots, l_p$ are called the loading factors for PC (standardized so: $\sum {l_i}^2 = 1$). They show how each X contributes to the PC
- PC's are uncorrelated: $\mathrm{Corr}(PC_i, PC_j) = 0$:
- PC_k are ordered so the first one (PC_1) explains most of the variance and so on

Example: Principal Component Analysis (TROPHY Data)

- Part of the metabolic risk score can be measured using the following10 variables:
  - Insulin
  - Glucose
  - Ins:Gluc ratio
  - Triglycerides
  - HDL
  - LDL
  - HDL:LDL ratio
  - Total Cholesterol
  - Systolic blood pressure
  - Diastolic blood pressure
- Each measure represents a health risk (cardiovascular risk). For HDL and HDL:LDL low score is bad, for the rest high score is bad

Example: Principal Component Analysis (TROPHY Data)

- So what happens when some scores are good and some are bad? We will use PCA to summarize the data in few meaningful PC's that still carry most of the information?
- Calculating principal components is easy (using R/SAS)
  - Interpreting what the components mean in scientific terms is not always easy

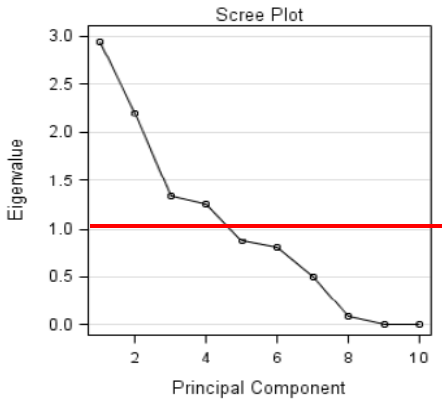
Example: Principal Component Analysis (TROPHY Data)

Output in R:

Importance of components:	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Proportion of Variance	.29	.22	.13	.12	.09	.08	.05	.01
Cumulative Proportion	.29	.51	.65	.77	.86	.94	.99	1.0

- PC1 alone explains 29% of the variance from the original data
- PC2 alone explains 22% of the variance from the original data
- Cumulative: PC1 and PC2 jointly explain 51% of the variance from the original data
- How to choose the number of PC?
  - Use Eigenvalues, Screen Plot

Screen Plot: Selecting the Number of PC's



- Eigenvalue criteria: Choose all PC's for which Eigenvalue > 1
- Visually: Look for the number of PC's where the curves start to flatten
- Explained Variance: Choose a small # of meaningful PC's that explain a "sufficient" amount of variance (e.g. 50-60%)

Interpretation of PC's: TROPHY Data

- Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Proportion of Variance	.29	.22	.13	.12	.09	.08	.05	.01
Cumulative Proportion	. 29	.51	.65	.77	.86	.94	.99	1.0

  - What is the interpretation of PC1, PC2, PC3?

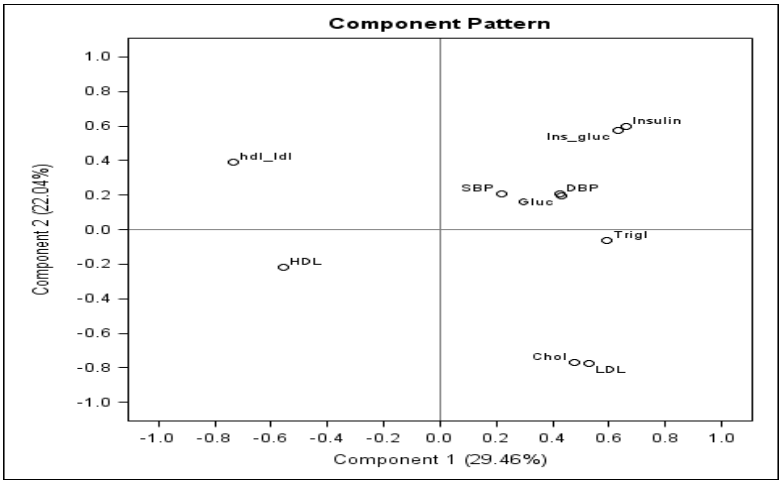
Interpretation of PC's: TROPHY Data

(Loading Factors)

	PC1	PC2	PC3
Insulin	0.38	0.40	-0.03
Gluc	0.25	0.13	-0.04
Ins_gluc	0.38	0.39	-0.05
Trigl	0.34	-0.04	-0.14
HDL	-0.32	-0.15	0.39
LDL	0.31	-0.52	0.09
hdl_ldl	-0.43	0.26	0.20
Chol	0.28	-0.52	0.17
SBP	0.13	0.14	0.56
DBP	0.25	0.14	0.66

PC1=.38Ins+.25Gluc+.38Ins_gluc+.34Trig-.32HDL+.3LDL-.43HDL_LDL
+.38Chol+.13SBP+.25DBP

Biplot of PC1 vs. PC2 in SAS



PCA Summary

(Example: TROPHY Data)

- The metabolic risk profile based on the 10 measures in the TROPHY example can be summarized using 2-3 PC's: PC1, PC2, PC3
  - PC1 is an overall weighted average score of the metabolic risk (high is bad)
  - PC2 is a contrast score: Insulin – Lipids
  - PC3 is an weighted average BP score (high is bad)
  - PC1 and PC2 explain 51% of the original variance
  - PC1,PC2,PC3 explain 65% of the original variance

Topic

- Multiple Testing
  - Family-Wise Error Rate
    - Bonferroni adjustment
  - False Discovery Rate
- Multivariate Data Analysis
  - Principle Component
    - Screen plot, Bipolar plot
  - **Cluster Analysis: K-means, Hierarchical Clustering**
    - **Dendrograms**
    - **Heatmaps**

Cluster Analysis

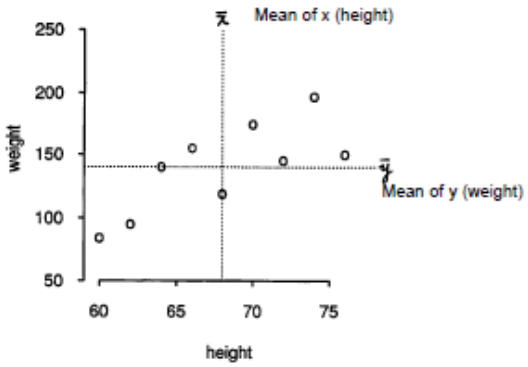
- Cluster analysis is a set of techniques that look for groups (clusters) in the data such that:
  - Individuals belonging to the same group resemble each other
  - Individuals belonging to different groups are dissimilar
- There are two main approaches of carrying out such allocation
  - **Partitional:** Partitioning into a number of clusters pre-specified by the user
    - K-means Method
  - **Agglomerative:** Starting with each individual as a separate cluster and aggregate similar individuals/clusters ending up with a single cluster of all individuals
    - Hierarchical Clustering

Example: Clustering Based on Two Variables

- Example data on height and weight for 9 people.

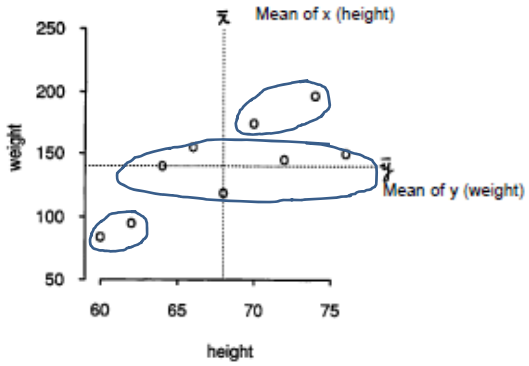
Height	Weight
60	84
62	95
64	140
66	155
68	119
70	175
72	145
74	197
76	150

Scatterplot: Plot of Height vs. Weight



Can you find 2 or 3 clusters?

Scatterplot: Plot of Height vs. Weight



Individuals must be closer within a cluster, but further between clusters.
How to measure being close? What type of distance to use?

Distance Measures for Cluster Analysis

- All clustering methods require the specification of a measure of "similarity". What individuals are considered similar (close) or dissimilar (far)?
- A **distance measure** is introduced to indicate distances between individuals, and subsequently between clusters
- Some **common** used distances are:
  - **Euclidian** or Square Euclidian
  - **Mahalanobis**
  - Maximum
  - Manhattan

Distance: How "Far" (Dissimilar) is X from Y

- For two subjects X and Y with data $x = (x_1, x_2, \dots, x_p)$ and y $= (y_1, y_2, \dots, y_p)$ the following distances can be used to measure the degree of similarity or dissimilarity:
  - **Euclidian distance**: $D(X,Y) = \sqrt{\sum_i (x_i - y_i)^2}$
  - **Mahalanobis distance**: $D_M(X,Y) = \sqrt{(x-y)^T \Sigma^{-1} (x-y)}$ (Σ is the covariance matrix)

Partitional Clustering: K-means Method

- Step 1: The **K-means** partitional clustering method starts with a random selection of K subjects for clusters $C_1, C_2, \dots, C_k$, where k is determined *a priori*
  - An initial cluster "center" is defined as $T_k = X_k$, for each cluster
- Step 2: Each subject is assigned to one of these clusters, based on the smallest distance from T_k ("center")
  - x is assigned to C_j if d(x, T_j) is the smallest
- Step 3: For the new clusters, calculate the new "centers" ($T_k = \overline{X}_k$) as the means of the subjects in each new cluster
- The procedure (step 2 and step 3) is repeated until no subjects are re-assigned

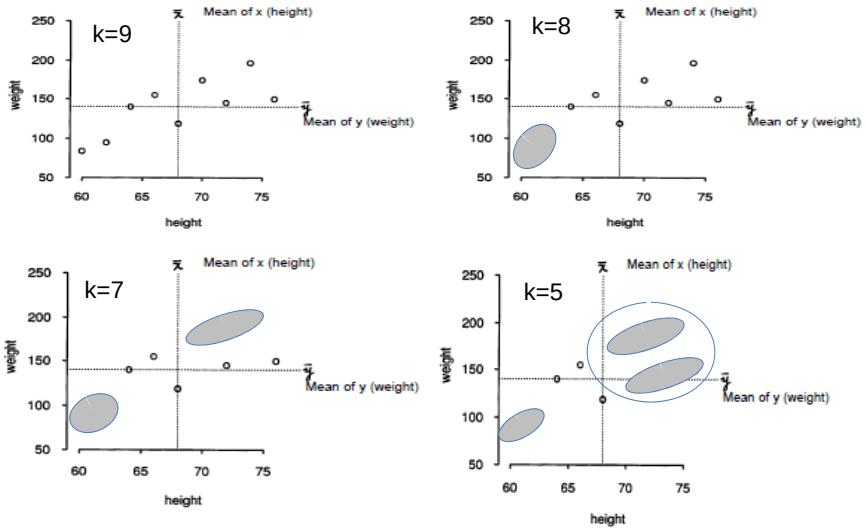
Partitional Clustering: K-means Method

- K-means is non-hierarchical clustering method. It is faster then hierarchical clustering
- It does not require specification of a linkage method (more on this later)
- The number of clusters, k, is pre-specified and fixed
- Hierarchical clustering, on the other hand, provides insight into the clustering process and does not required a pre-specified number of clusters

Hierarchical Clustering

- All hierarchical clustering methods start with each individual defining its own cluster. Then clusters are joined sequentially in a hierarchical way
- How are two clusters joined:
  - Calculate the distance, D(C_i, C_j), between every pair of clusters based on one a linkage criteria (more later)
  - Then join the two "nearest" clusters who have the smallest D(C_i, C_j)
  - Continue until there is only one cluster

Example: Hierarchical Clustering for Height/Weight Data

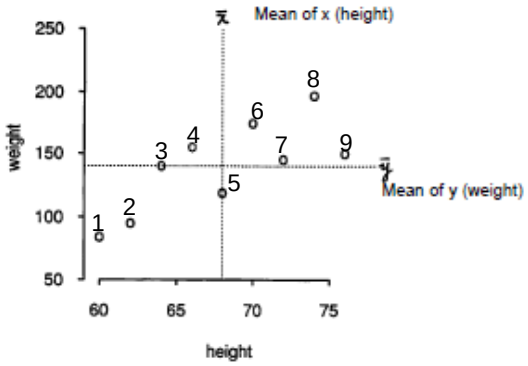


In 2D and with small data this is possible. With more than 2 variables Dendrogram, is a better way

Dendograms

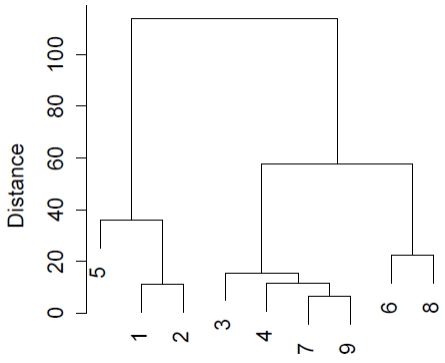
- Dendrogram is a useful graphical tool for displaying multidimensional hierarchical structure of clustering
- It shows the distances between individuals (and clusters) in a tree-like structure
- Individuals (or clusters of individuals) that are closest to each other are connected by a horizontal line, forming a new cluster

Dendrogram Example for Height Weight Data



Dendrogram Example for Height Weight Data

- Merging clusters is based on a linkage criteria:
  - Single linkage
  - Complete linkage
  - Average linkage
  - Ward linkage

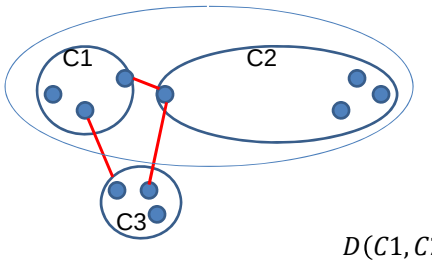


The distance of a particular pair of objects (or clusters) is reflected in the height of the horizontal line. It is based on the linkage criteria

Single Linkage Clustering (Minimum)

In single linkage, the distance between two clusters is computed as the distance between the two closest elements in the two clusters:

$$D(C1, C2) = \min_{x \in C1; y \in C2} \{d(x, y)\}$$

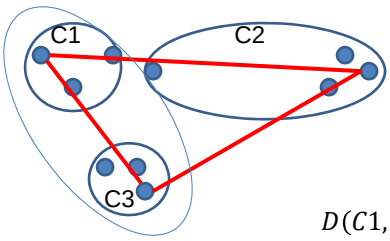


$D(C1, C2)$ is the shortest distance among C1, C2, and C3, so link C1 and C2

Complete Linkage Clustering (Maximum)

In complete linkage, the distance between two clusters is computed as the distance between the two farthest elements in the two clusters:

$$D(C1, C2) = \max_{x \in C1; y \in C2} \{d(x, y)\}$$

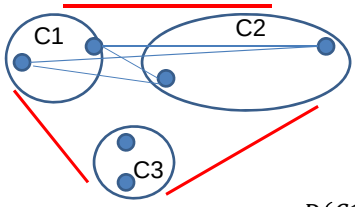


$D(C1, C3)$ is the shortest distance among C1, C2, and C3, so link C1 and C3

Average Linkage Clustering (Mean)

In average linkage, the distance between two clusters is computed as the mean of all distances between pairs of elements in the two clusters

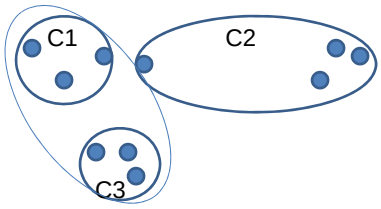
$$D(C1, C2) = \underset{x \epsilon C1, y \epsilon C2}{mean} \{d(x, y)\}$$



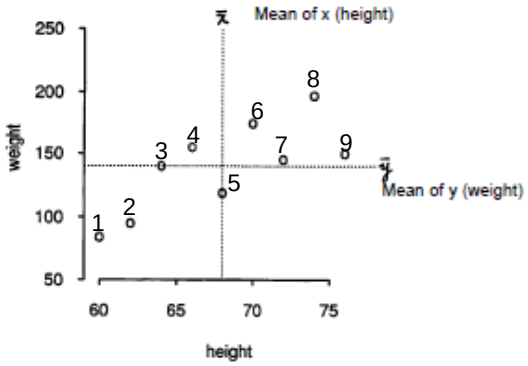
$D(C1, C3)$ is the shortest distance among C1, C2, and C3, so link C1 and C3

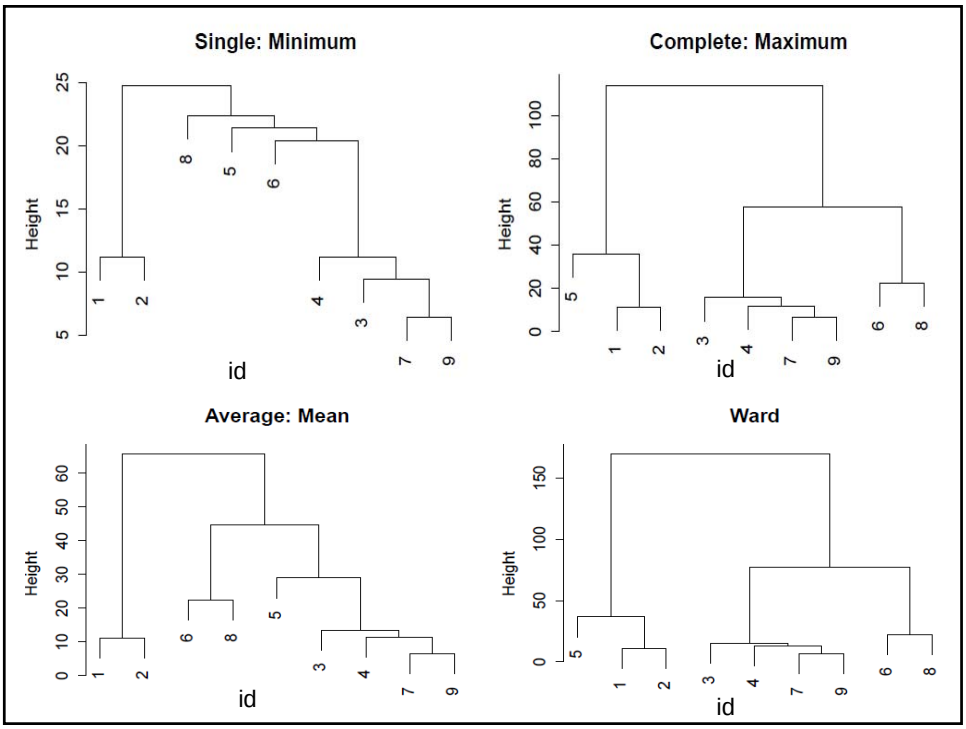
Ward Linkage Clustering

Ward's criterion minimizes the total within-cluster variance. At each step the pair of clusters that result in a minimum increase in variance are merged



Dendrogram Example for Height Weight Data Using Different Linkage Rule

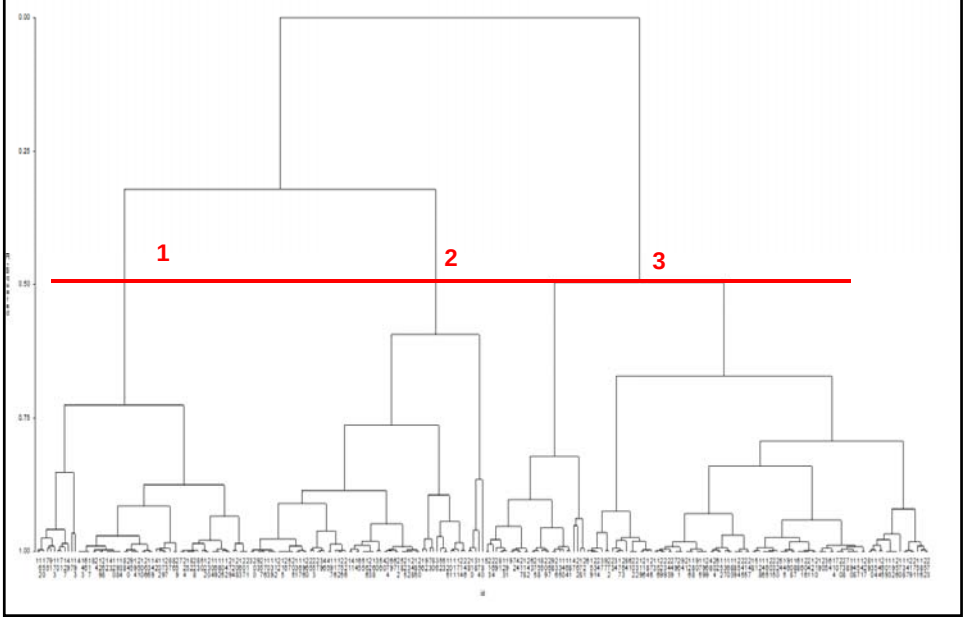




Illustrating Cluster Analysis Using TROPHY Data

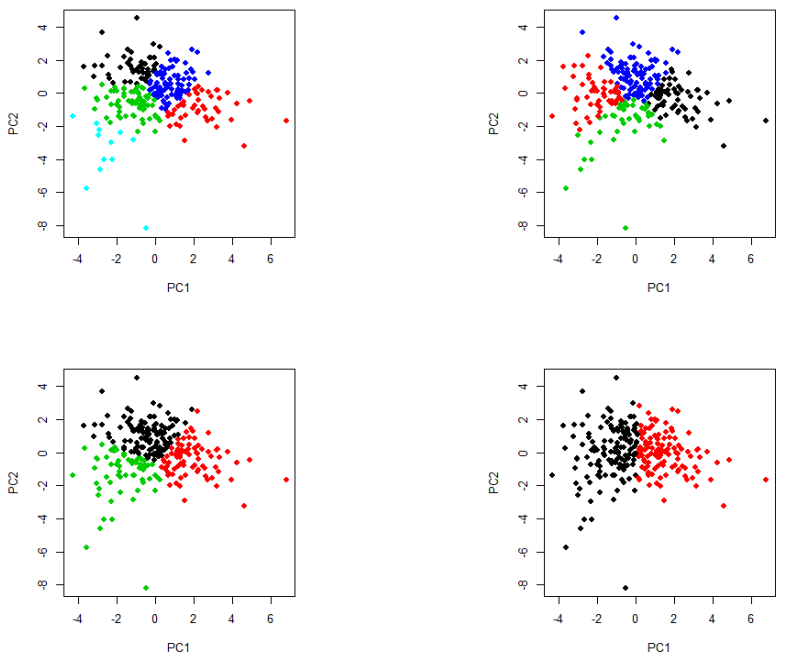
- We will use TROPHY data to group subjects into clusters based on their metabolic risk measures
  - Insulin, Glucose, Ins:Gluc ratio, Triglycerides, HDL, LDL, HDL:LDL ratio, Total Cholesterol, Systolic blood pressure, and diastolic blood pressure
- Earlier in PCA we showed that PC1 and PC2 contain most of the information on the metabolic risk
- So, it will be simpler to run cluster analysis based on PC1 and PC2 alone, without losing much of the information of the original data

Dendrogram for TROPHY Example



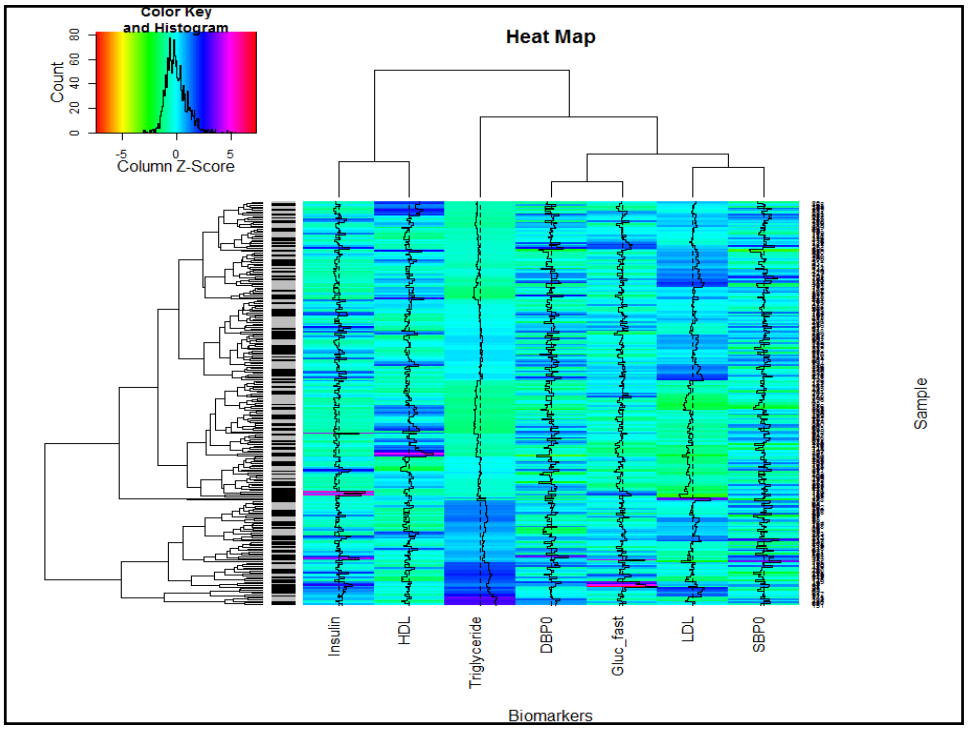
Naming Clusters

- Becouse cluster analysis is an unsupervised process for identifying clusters, giving a name to each cluster it's not easy
  - Usually, you look for specific features (based on X1,X2,…,Xp measures) to give an appropriate name
  - This may be hard, if clustering is based on a large number of variables
  - The data reduction via PCA make it easier, as its focused on features specific to few PC's (i.e. PC1, PC2)



Heatmap

- A **heatmap** is a graphical representation of multidimensional structure data using colors
  - In a heatmap individual values are distinguished by colors. E.g. large values are colored red, low values yellow
  - Dendogram is often added to a heatmap by permuting the rows/columns of a matrix to place similar values near each other
  - Examples: DNA microarrays data. Represent the level of expression of many genes across a number of comparable samples



Summary Points

- Adjustment in Multiple Testing (e.g. testing *m* hypothesis)
  - Family-Wise Error Rate (FWER)
    - Bonferroni adjustment (works for small m)
    - Set the significance of each test at $p_k \leq \frac{\alpha}{m}$. Then FWER $\leq m(\frac{\alpha}{m}) = \alpha$
  - False Discovery Rate (FDR)
    - Control the FDR $\leq \alpha$ (α is an excepted rate of false discoveries)
    - For each of m tests get the p-value. Order them: $p_1 \leq p_2 \leq \ldots \leq p_m$. Find largest k, such that $p_k \leq \frac{k*.05}{m}$. Then $p_1, p_2, \ldots, p_k$ are consider significant (reject $H_{01}, \ldots, H_{0k}$)

Summary Points

- ***Principal component analysis*** (PCA) is a dimension-reduction technique that looks for structure among variables $(X_1, X_2, \ldots, X_p)$
- PCA finds a small number of uncorrelated **linear combinations** of the variables (X's), which summarize most of the information from X's
  - Number of PC's. Select the number of PC's based on:
    - Eigenvalue criteria: Chose all PC's for which Eigenvalue > 1
    - Visually: Use the screen plot to identify the number of PC's where the plot start to flatten
    - Explained Variance: Chose a # of PC's that explain a "sufficient" amount of the variance
  - Interpretation of PC's. Use biplots to see how each of the original data (X's) contributes to a PC

Summary Points

- **Cluster Analysis** identifies clusters of individuals/objects in a dataset that are similar based on a distance (e.g. Euclidian, Mahalanobis)
  - Partition (non-hierarchical method)
    - Use K-mean method to find k (pre-specified) clusters
  - Hierarchical clustering
    - Identify clusters staring with each individual as its own cluster. Next merge clusters (using a linkage criteria) hierarchically until all are part of one cluster
    - Common linkage criteria:

      Single linkage, complete linkage, average linkage, Ward linkage
  - Dendrograms: A visual tree-like structure describing the hierarchical nature of clustering in the data
- **Heatmap:** A graphical representation of multidimensional structure data using colors