Measures of Disease and Unsupervised Learning Methods

Alexander McLain

June 6, 2024

Risk definition

- **Risk**- the probability that an individual develops a given disease over a specified period, given they are at risk for the disease.
- Risk can be crudely estimated as:

$$Risk = \frac{\#\text{ of new cases in a time period}}{\#\text{ at risk in the begining of follow-up}}$$

- Prevalence is generally estimated as a risk (probability).
- Risk is a probability statement (0¡Risk¡1) assuming an individual is not removed for any other reason during a given period of time.

Examples

- Point prevalence is the risk of having condition at a single point in time.
- Period prevalence is the risk of having condition for any part of a time interval or period.
- Survey all classes on a given day to estimate the point prevalence of colds.
- Survey all classes in May to ask how many had the flu during the past 12 months to estimate period prevalence.
- 1 in a million chance of developing cancer in a 70 year lifetime
- Problems with the crude estimate:
  - Generally, the population is dynamic
  - Competing risks: what if person dies from another cause?
  - Different follow-up periods are often part of study design

Rate definition

- **Rate** – instantaneous potential for change in some quantity (e.g., the number of incident cases) per unit change in another quantity (e.g., time)
- Usually report as rate per some number of people:

$$Rate = \frac{\#\text{ of new cases}}{\text{person-time units of follow-up}}$$

- Rates tell us how fast the disease is occurring in a population (speed is a good way to think of it).
  - For example, 70 new cases of breast cancer per 1,000 women per year (sometimes leave off 'per year').
- The interpretation of "rates" are more difficult then for "risk."

Rate versus risk

- Rates can be used to estimate risk if the time period is short (annual) and the incidence of disease over the interval is relatively constant (i.e., constant hazard).
- This is sometimes represented in the mathematical formulae

$$Risk = Rate \times Time,$$

$$Rate = \frac{Risk}{\Delta},$$

or

$$Prevalence = Incidence \times Duration$$

Rate to risk example

- Suppose that we have a population of 1000 persons in which the **incidence rate of cancer X** is 6 cases per 1000 person-years (6 / 1000 yr^{-1}).
- If we follow this population for 30 years the **risk of cancer X** in the population over that 30 years is: 6 / 1000 yr^{-1} x 30 years = 0.18 or 18%.
- Among the 1000 persons present at the start of the follow-up, 180 cases of cancer X will occur.
- If the follow up was 15 years the related risk would be 9%.
- The above formula does not take into account the decrease of the population at risk over time and cannot be used when risk is large. It also assumes that rate remains constant over time.

Example

Example 1:

Incidence of coronary heart disease observed in 12 years of follow-up (Framingham heart study; Schlesselman, p. 29)

Age	# at risk	# CHD	Risk	P-Years	Inc. Rate per 1000	Calc. Risk
30-39	789	40	0.050697	9228	4.335	0.05069
40-49	742	88	0.118598	8376	10.506	0.11845
50-59	656	130	0.198171	7092	18.331	0.19746

For ages 30-39:

Crude risk: 0.050597 = 40/789 (i.e., simple probability)

Incidence rate: 4.335 = 40/9228 *1000 (average incidence)

Risk: 0.05069 = 1 - exp(-0.004335*12) (12 years)

Risk and Rate Ratio

- Often we will use ratio measures of association, either risk ratio or rate ratio.
- In practice, choice of risk ratio or rate ratio makes little difference, as long as the following assumptions are met:
  - low background rate
  - short follow-up
  - weak association between exposure and disease
- Commonly we'll want to compare rates or risks across groups, say between exposed and non-exposed groups.
- Often the relative difference (i.e., the risk for the exposed relative to the non-exposed) is of more importance than the absolute difference.

Example

Example 2: for illustration of basic measures of association

Approximate number of new cases and incident rates of lung cancer and coronary heart disease for the US population in 1976, by smoking status.

		Lung Cancer		Coronary Heart Disease	
	Population	N (cases)	Rate/100,000	N (cases)	Rate/100,000
Smoker	70,000,000	60,000	85.7	250,000	357.1
Non-smoker	150,000,000	10,000	6.7	250,000	166.7
Total	220,000,000	70,000	31.8	500,000	227.3

Excess risk and attributable risk

- Let λ_1 and λ_0 denote the incident rates for the "exposed" and "non-exposed" groups, respectively.
- The **excess risk** is then

$$b = \lambda_1 - \lambda_0$$

- Interpretation: *The rate of disease in the exposed group is b greater than the rate among non- exposed.*

Excess risk and attributable risk

- Let λ_1 and λ_0 denote the incident rates for the "exposed" and "non-exposed" groups, respectively.
- The **excess risk** is then

$$b = \lambda_1 - \lambda_0$$

- Interpretation: *The rate of disease in the exposed group is b greater than the rate among non- exposed.*
- The **attributable risk among exposed persons** is

$$AR = \frac{\lambda_1 - \lambda_0}{\lambda_1}$$

- Interpretation: *Among the exposed the proportion (or %) with the disease that can be attributed to the exposure.*

Population Attributable Risk

- Let p =the proportion of the population that is exposed. (How could we estimate this?)
- Then the total disease incidence rate is

$$\lambda = p\lambda_1 + (1-p)\lambda_0 = \lambda_0 + p(\lambda_1 - \lambda_0)$$

- The population attributable risk (PAR) is

$$PAR = \frac{\lambda - \lambda_0}{\lambda} = \frac{p(\lambda_1 - \lambda_0)}{p\lambda_1 + (1-p)\lambda_0}$$

- Interpretation: *The population attributable risk is the proportion of the disease occurrence in the total population that can be attributed to the exposure, i.e., that is beyond the baseline risk.*

Multiplicative Model

- Another way to express the increase (or decrease) in risk/rate with respect to an exposure is to use a multiplicative model.
- In this case we are interested in
$$r = \frac{\lambda_1}{\lambda_0}$$
- Interpretation: *The rate of disease in the exposed group is r times higher than the rate of disease in the non-exposed group.*
- Relative risk
$$RR = \frac{\Pr(D|exposed)}{\Pr(D|\text{non-exposed})}$$
the numerator and denominator are commonly estimated from the cumulative hazard rate $\Lambda(t) \approx R(t)$.

Which to use?

This will depend on your question.

- If you are looking at the total disease burden impact on the population, PAR is probably relevant.
- If you are looking at the potential impact of an intervention, e.g., to reduce smoking, the AR among exposed may be more relevant.
- The multiplicative model has appealing properties in terms of statistical estimation and theory.
  - It does not require the estimation of both rates.

Properties of relative risk

- Relative risk has many appealing properties.
- The relative risk is only estimable from cohort studies.
- The odds ratio approximates the relative risk;
- The odds ratio can be estimated from a cohort or case-control study.

Estimation of OR from a case-control study

- p_e =proportion of population exposed to study factor ($q_e = 1 - p_e$).
- $R_1 = \Pr(D|exposed)$, $R_0 = \Pr(D|not\ exposed)$, $Q_1 = 1 - R_1$, and $Q_0 = 1 - R_0$.
- Then if R_1 and R_0 are small, such that $1 - R_1 \approx 1$ and $1 - R_0 \approx 1$, the risk ratio and rate ratio are approximately the same

$$OR = \frac{\frac{R_1}{1-R_1}}{\frac{R_0}{1-R_0}} = \frac{R_1(1-R_0)}{R_0(1-R_1)} \approx \frac{R_1}{R_0} = r$$

- Example:
  - $R_1 = 0.002$ and $R_0 = 0.001$
  - $R_1 = 0.2$ and $R_0 = 0.1$

2 x 2 Table

The general format of a 2x2 table is:

	Diseased	Not Diseased	Totals
Exposed	a	b	m_1
Not Exposed	c	d	m_0
Totals	n_1	n_0	N

- Mostly, the outcome status is used as the column variable.
- Not important for estimating OR, but important for interpreting the results and estimating relative risk

Relative Questions

Questions:

- Is disease prevalence the same in exposed and non-exposed groups?
  - Appropriate for cohort studies.
- Is exposure prevalence the same in diseased and non-diseased groups?
  - Appropriate for case-control studies.
- To answer either of these questions, we need to compare 2 proportions, which requires estimates of the two risks/rates.

Basic Properties

- From the 2×2 we usually we'll estimate the OR using

$$\widetilde{OR} = \frac{ad}{bc}$$

- We can do hypothesis tests of $H_0 : OR = 1$ from the 2×2 table.

Introduction

- Rarely is it reasonable to talk about the association of two variables without consideration of the impact of other variables.
- You will see a variety of labels for different etiologic and statistical phenomena that might be occurring.
- For example
  - Confounding
  - Direct Effect.
  - Indirect Effect.
  - Mediation
  - Modification/moderation
  - Interaction

Confounding

- Confounding reflects the causal association between variables in the population under study
- A confounder is an extraneous variable that satisfies the following criteria:
  - It is a risk factor for the study disease.
  - It is associated with the study exposure, but is not a consequence of that exposure.
  - The association with disease must occur in the absence of exposure.
- A confounder is a risk factor for the study disease whose "control" in some appropriate way will reduce (or remove) bias in estimating the exposure–disease relationship.

Confounding

- If we are interested in estimating the OR then we could get a biased estimate of the OR if a confounder is not adjusted for (e.g., looking at OC–MI, while not adjusting for age/smoke).
  - OR could be positively or negatively biased.
- How do we control/identify confounding?
  - Must have prior knowledge on potential confouders
  - Is the adjusted estimate of the OR $\approx$ to the crude estimate of the OR?
  - People often use the 10% rule (i.e., a 10% change in the OR is a sign of confounding).

Berkley Gender Bias Case:

- Data from: "Sex Bias in Graduate Admissions: Data from Berkeley," *Science* 187: 398-403; 1975.
- In 1973 2,681 men and 1,835 women applied to graduate school, with 44% of men and 35% of women being admitted.
- A crude analysis finds Crude OR $= (1276/1835)/(1486/2681) = 1.25$ with 95% CI $(1.20, 1.32)$.
- This difference is statistically significant (i.e., not due to chance).

Berkley Gender Bias Case:

The admission rates and RR by department

	Men		Women		
Depart.	Applicants	Admitted	Applicants	Admitted	OR_i
A	825	62%	108	82%	0.90
B	560	63%	25	68%	0.99
C	325	37%	593	34%	1.08
D	407	35%	375	34%	1.02
E	191	23%	393	26%	0.88
F	373	7%	341	6%	1.09

- The Maentel-Haenszel Summary OR is 0.97 and insignificant.
- The apparent association (OR=1.25) was due to confounding.

Mediation

- Can be thought of as a special case of a potential confounder being part of a causal pathway.
- if $E \Rightarrow M \Rightarrow D$ then we say M is a mediator.

For example:

- TAAG (Trial for Activity among Adolescent Girls)
- Interventions (E) effected physical activities (D).
- The E may have changed self-image or self efficacy (M).
- The could be a relationship between M and D.
- Important to understand why physical activity increased.

Introduction

- When we observe that the OR are non-constant across strata, we say there is **statistical interaction**.
- Interaction (a statistical term) is related to effect modification (an epi term).
- If the effect of E on D varies with C we say that there C is an effect modifier.
  - **Effect modification:** a variable that differentially (positively and negatively) modifies the observed effect of a risk factor on disease status. Present if different groups (of C) have different risk estimates.
  - Effect modification is related to the biology of disease, not just a data observation (what makes it different from interaction).
- An obvious example of interaction is breast or prostate cancer.

Introduction

Why study effect modification? Why do we care?

- to define high-risk subgroups for preventive actions,
- to increase precision of effect estimation by taking into account groups that may be affected differently,
- to increase the ability to compare across studies that have different proportions of effect-modifying groups, and
- to aid in developing a causal hypotheses for the disease.

Effect Modification vs Confounding

Interesting tidbit:

- If a variable C is a confounder, then the stratum specific estimates of the $\widehat{OR}_i$ will mostly be on one side of the crude estimate of the $\widehat{OR}$.
- If a variable C is an effect modifier, then the crude estimate of the $\widehat{OR}$ will be a "weighted average" of the stratum specific estimates of the $\widehat{OR}_i$.
- **Note:** we are comparing the crude OR **NOT** the adjusted estimate of the OR.

What to do under effect modification/confounding?

1. If a variable is a confounder, you just have to adjust for it:

$$\beta_0 + \beta_1 E + +\beta_2 C$$

2. If a variable is an effect modifier, it should go into your regression equation as an **interaction term**

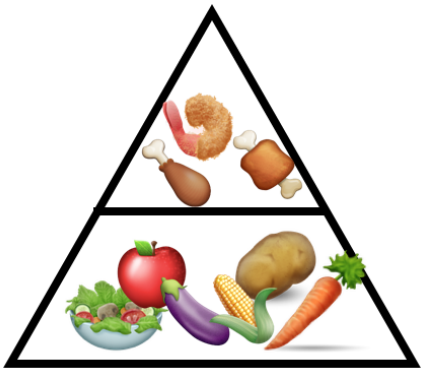
$$\beta_0 + \beta_1 E + +\beta_2 C + \beta_3 EC$$

Introduction

- Principal components explain the variance-covariance structure of a set of variables through some linear combinations.
- Knowing the variables that best differentiate your items has several uses:
  1. Visualization. Using the right variables to plot items will give more insights.
  2. Uncovering Clusters. With good visualizations, hidden categories or clusters could be identified.

Introduction

- Consider data from the United States Department of Agriculture.
- This data contains the nutritional content of one serving of several food items (Parsley, Kale, Broccoli, Corn, Chick, Beef, etc.).
- Four nutrition variables were analyzed: Vitamin C, Fiber, Fat and Protein.
- We want to consider making one variable that fully characterizes the food.




Vitamin C
Parsley
Kale
Broccoli
Cauliflower
Soybeans
Yam
Guinea Hen
(Vitamin C) - (Fat)
Parsley
Kale
Broccoli
Cauliflower
Cabbage
Spinach
Yam
Sweet Corn
Guinea Hen
Bluefish
Mackerel
Chicken
Beef
Pork
Lamb
(Vitamin C + Fiber) - (Fat)
Parsley
Peas
Lotus Root
Chives
Cauliflower
Soybeans
Eggplant
Sweet Corn
Mushrooms
Haddock
Guinea Hen
Bluefish
Mackerel
Chicken
Beef
Pork
Lamb

Introduction

- Principal components explain the variance-covariance structure of a set of variables through some linear combinations.
- $\boldsymbol{X}$: variables which we want to explain the variance-covariance matrix
- $\boldsymbol{Z}$: a linear combination of the X's (this is not the response)

$$
\begin{aligned}
\boldsymbol{Z}_1 &= \boldsymbol{X}\boldsymbol{v}_1 = v_{11}\boldsymbol{X}_1 + v_{12}\boldsymbol{X}_2 + \ldots + v_{1p}\boldsymbol{X}_p \\
\boldsymbol{Z}_2 &= \boldsymbol{X}\boldsymbol{v}_1 = v_{21}\boldsymbol{X}_1 + v_{22}\boldsymbol{X}_2 + \ldots + v_{2p}\boldsymbol{X}_p \\
\vdots \quad & \quad \vdots \quad \vdots \\
\boldsymbol{Z}_p &= \boldsymbol{X}\boldsymbol{v}_1 = v_{p1}\boldsymbol{X}_1 + v_{p2}\boldsymbol{X}_2 + \ldots + v_{pp}\boldsymbol{X}_p
\end{aligned}
$$

PC goal

Main idea of Principal Components

- Set $\boldsymbol{Z}_1 = \boldsymbol{X}\boldsymbol{v}_1$ such that the variance of $\boldsymbol{Z}_1$ is maximized over all $\boldsymbol{v}$ such that $\boldsymbol{v}'\boldsymbol{v} = 1$.
- Set $\boldsymbol{Z}_2 = \boldsymbol{X}\boldsymbol{v}_2$ such that the variance of $\boldsymbol{Z}_2$ is maximized over all $\boldsymbol{v}$ such that $\boldsymbol{v}'\boldsymbol{v} = 1$ and $Cov(\boldsymbol{Z}_1, \boldsymbol{Z}_2) = 0$.

⋮

- Set $\boldsymbol{Z}_i = \boldsymbol{X}\boldsymbol{v}_i$ such that the variance of $\boldsymbol{Z}_i$ is maximized over all $\boldsymbol{v}$ such that $\boldsymbol{v}'\boldsymbol{v} = 1$ and $Cov(\boldsymbol{Z}_k, \boldsymbol{Z}_i) = 0$ for all $k = 1, 2, \ldots, i = 1$.

⋮

These $\boldsymbol{v}$ values can be found using some linear algebra techniques.

	PC1	PC2	PC3	PC4
Fat	-0.45	0.66	0.58	0.18
Protein	-0.55	0.21	-0.46	-0.67
Fiber	0.55	0.19	0.43	-0.69
Vitamin C	0.44	0.70	-0.52	0.22

Figure: Factor loadings for the food example.

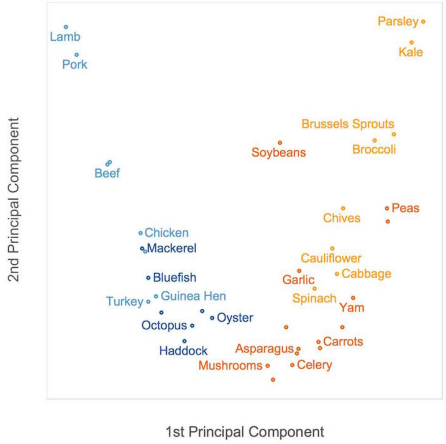


Figure: First and second PCs for the food example.

Properties

- Let d_k denote the variance of Z_k
- The total proportion variability that can be attributed to Z_k is

$$\frac{d_k}{\sum_{i=1}^{p} d_i}$$

- Suppose

$$\frac{d_1 + d_2 + d_3}{\sum_{i=1}^{p} d_i} = 0.97.$$

what would this imply about $Z_4, \ldots, Z_p$?

How many PC's to use

- If $\frac{\sum_{i=1}^{r} d_i}{\sum_{i=1}^{p} d_i}$ is "close to one," then only the first r PC's are needed.
- Scree plot: a plot of d_i used for selecting r
- *Kaiser's Rule* for the PC of a correlation matrix cutoff at 1.
- PC's are not invariant to scaling

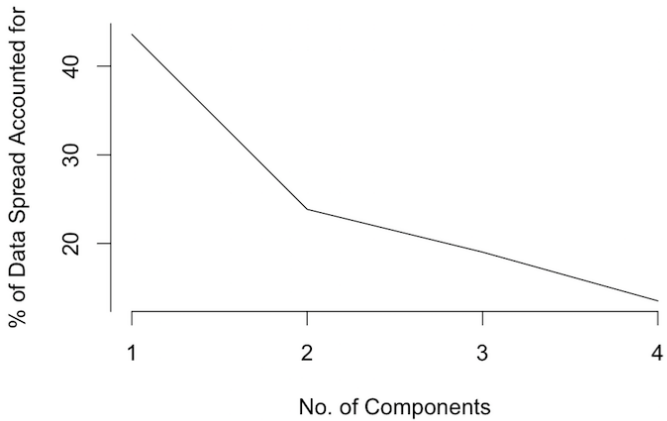


Figure: The total proportion variability attributed to the PCs.

Principal components regression

- Principal components regression (PCR) forms the linear combination $\boldsymbol{z}_m = \boldsymbol{X}\boldsymbol{v}_m$ and then regresses $\boldsymbol{y}$ on $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_M$ for some $M \leq p$.

K-means algorithm

- The term "k-means" was first used by James MacQueen in 1967, though the idea goes back to Hugo Steinhaus in 1957.
- The standard algorithm was first proposed by Stuart Lloyd in 1957 as a technique for pulse-code modulation, though it wasn't published outside of Bell Labs until 1982.
- In 1965, E. W. Forgy published essentially the same method, which is why it is sometimes referred to as Lloyd-Forgy.

Dissimilarity

- We first need to choose a dissimilarity (or distance) metric to use.
- The dissimilarity between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ has the following properties
  1. $d(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0$
  2. $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0 \leftrightarrow \boldsymbol{x}_i = \boldsymbol{x}_j$
  3. $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = d(\boldsymbol{x}_j, \boldsymbol{x}_i)$

Dissimilarity

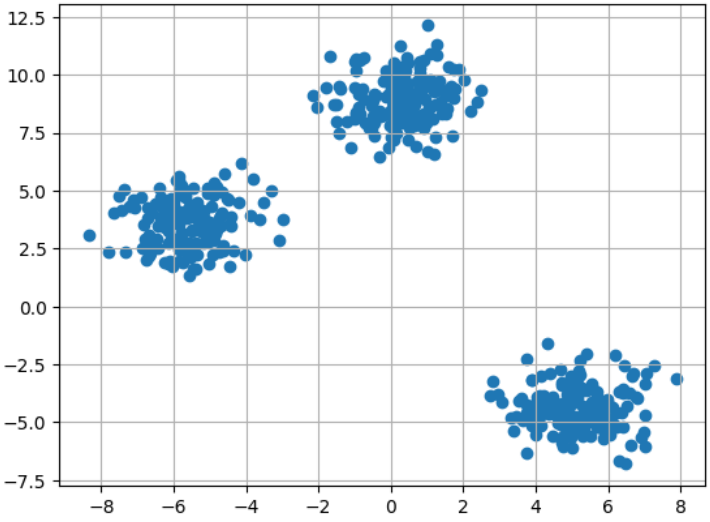
- We first need to choose a dissimilarity (or distance) metric to use.
- The dissimilarity between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ has the following properties
  1. $d(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0$
  2. $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0 \leftrightarrow \boldsymbol{x}_i = \boldsymbol{x}_j$
  3. $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = d(\boldsymbol{x}_j, \boldsymbol{x}_i)$
- **Examples**
  - Euclidean: $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)'(\boldsymbol{x}_i - \boldsymbol{x}_j)}$
  - Manhattan: $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$
  - $1-$correlation: $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 - \rho_{ij}$
- Let $\boldsymbol{D} = (d_{ij})_{i,j,=1,\ldots,n}$ where $d_{ij} = d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ be the *proximity matrix*

K-means Algorithm

- To perform K-means, we first chose K cluster centers $\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots, \boldsymbol{m}_K$. These can simply be K observations in the data choosen at random.
- Algorithm:
  1. Group each observation to the closest cluster center in the choosen dissimilarity metric. That is, observation $\boldsymbol{x}_i$ would be put in cluster k if

$$d(\boldsymbol{x}_i, \boldsymbol{m}_k) < d(\boldsymbol{x}_i, \boldsymbol{m}_j) \text{ for all } j \neq k$$

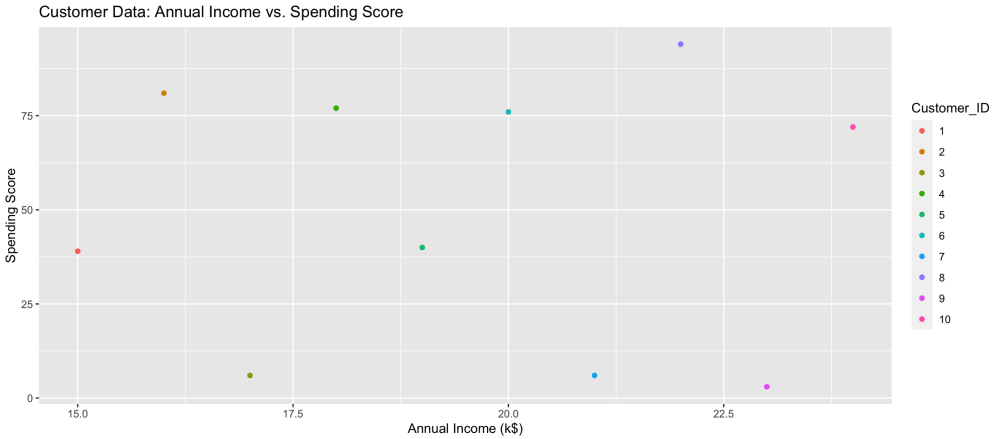
  2. Re-calculate the the K centers. $\boldsymbol{m}_k$ is set to the mean or median of all $\boldsymbol{x}_i$ in cluster k.
- Steps 1 and 2 are iterated until convergence (i.e., the cluster centers stop moving or the groups stop changing).

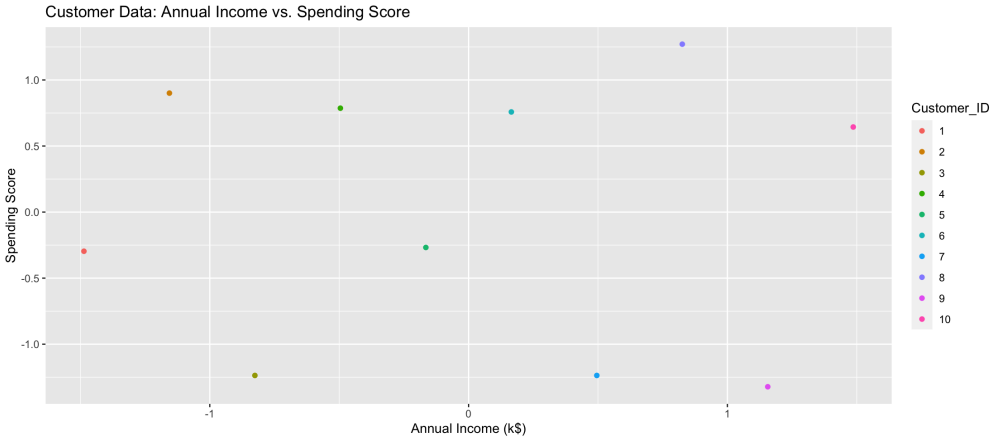

12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−8
−6
−4
−2
0
2
4
6
8

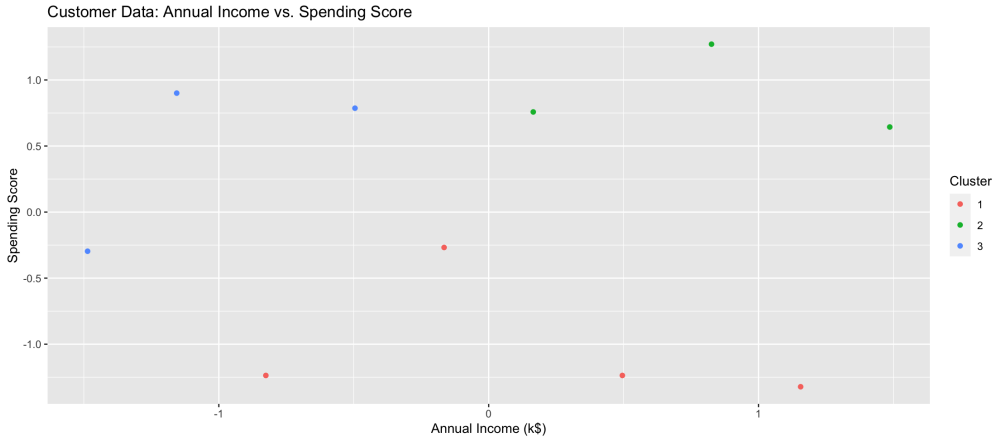
K-means Example

Customer ID	Annual Income	Spending Score
1	15	39
2	16	81
3	17	6
4	18	77
5	19	40
6	20	76
7	21	6
8	22	94
9	23	3
10	24	72

Let's start with observations 3, 4, and 10 as the "centers"


Customer Data: Annual Income vs. Spending Score
Spending Score
75
50
25
0
15.0
17.5
20.0
22.5
Annual Income (k$)
Customer_ID
1
2
3
4
5
6
7
8
9
10


Customer Data: Annual Income vs. Spending Score
Spending Score
1.0
0.5
0.0
-0.5
-1.0
-1
0
1
Annual Income (k$)
Customer_ID
1
2
3
4
5
6
7
8
9
10


Customer Data: Annual Income vs. Spending Score
Spending Score
1.0
0.5
0.0
-0.5
-1.0
-1
0
1
Annual Income (k$)
Cluster
1
2
3

Hierarchical Clustering

Algorithm for agglomerative clustering:

1. Let $\mathcal{L} = \{\boldsymbol{x}_i; i = 1, 2, \ldots, n\}$ be the n-clusters
2. Find the smallest element of $\boldsymbol{D} = d_{IJ}$. Put I and J in the same cluster.

Now we have to recompute the dissimilarities for the new group IJ.

3. Compute dissimiarities $d_{IJ,k}$ for all $k \neq I$ or J. Three options for doing this:
   - **single**: $d_{IJ,k} = \min(d_{Ik}, d_{Jk})$
   - **complete**: $d_{IJ,k} = \max(d_{Ik}, d_{Jk})$
   - **average**: $d_{IJ,k} = avg(d_{Ik}, d_{Jk})$
4. For $\boldsymbol{D}^{(1)}$ which removes row I and J and includes the row calculated in 3.
5. Repeat steps 2–4 $n - 1$ times.

Hierarchical Clustering

- Note that $\min(d_{Ik}, d_{Jk}) < avg(d_{Ik}, d_{Jk}) < \max(d_{Ik}, d_{Jk})$
- Single is more likely to cluster an observation with a group of observations than complete (average is in the middle). There are other options (e.g., median).

- The algorithm for divisive clustering begins with the entire data set as a single cluster, and recursively divides one of the existing clusters into two daughter clusters at each iteration in a top-down fashion.
- How to divide is not straightforward, and divisive clustering is not as common as agglomerative.

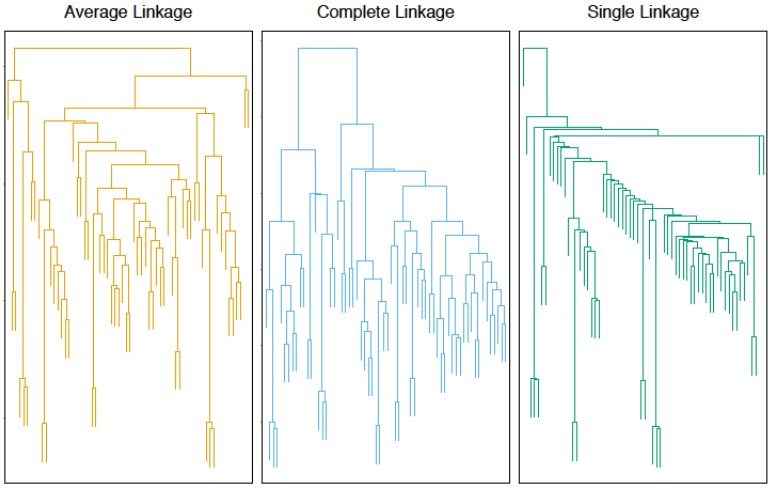


Figure: From page 524 in the online version of ESL.

Example

	1	2	3	4	5	6	7	8	9
2	42.01								
3	33.06	75.01							
4	38.12	4.47	71.01						
5	4.12	41.11	34.06	37.01					
6	37.34	6.4	70.06	2.24	36.01				
7	33.54	75.17	4	71.06	34.06	70.01			
8	55.44	14.32	88.14	17.46	54.08	18.11	88.01		
9	36.88	78.31	6.71	74.17	37.22	73.06	3.61	91.01	
10	34.21	12.04	66.37	7.81	32.39	5.66	66.07	22.09	69.01

Table: Dissemilarity matrix for example data

Hierarchical Example

Customer ID	Annual Income	Spending Score
1	15	39
2	16	81
3	17	6
4	18	77
5	19	40
6	20	76
7	21	6
8	22	94
9	23	3
10	24	72

Example

	1	2	3	4/6	5	7	8	9
2	42.01							
3	33.06	75.01						
4/6	37.73	5.44	70.54					
5	4.12	41.11	34.06	37.01				
7	33.54	75.17	4	71.06	34.06			
8	55.44	14.32	88.14	17.46	54.08	88.01		
9	36.88	78.31	6.71	74.17	37.22	3.61	91.01	
10	34.21	12.04	66.37	7.81	32.39	66.07	22.09	69.01

Table: Dissemilarity matrix for example data after step 1 using the average method.

Example

	1	2	3	4/6	5	7/9	8
2	42.01						
3	33.06	75.01					
4/6	37.73	5.44	70.54				
5	4.12	41.11	34.06	37.01			
7/9	35.21	76.74	5.35	72.62	35.64		
8	55.44	14.32	88.14	17.46	54.08	88.01	
10	34.21	12.04	66.37	7.81	32.39	66.07	22.09

Table: Dissemilarity matrix for example data after step 2 using the average method.

Example

	1/5	2	3	4/6	7/9	8
2	42.01					
3	33.06	75.01				
4/6	37.73	5.44	70.54			
7/9	35.21	76.74	5.35	72.62		
8	55.44	14.32	88.14	17.46	88.01	
10	34.21	12.04	66.37	7.81	66.07	22.09

Table: Dissemilarity matrix for example data after step 3 using the average method.

Example

	1/5	2	3/7/9	4/6	8
2	42.01				
3/7/9	34.14	75.87			
4/6	37.73	5.44	70.54		
8	55.44	14.32	88.14	17.46	
10	34.21	12.04	66.37	7.81	22.09

Table: Dissemilarity matrix for example data after step 4 using the average method.

Example

	1/5	2/4/6	3/7/9	8
2/4/6	39.87			
3/7/9	34.14	75.87		
8	55.44	14.32	88.14	
10	34.21	12.04	66.37	22.09

Table: Dissemilarity matrix for example data after step 5 using the average method.

Example

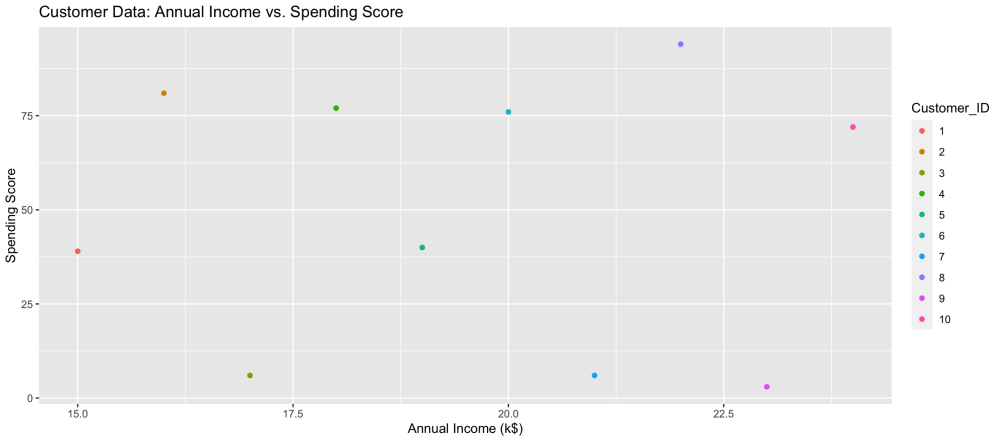
	1/5	2/4/6/10	3/7/9
2/4/6/10	37.04		
3/7/9	34.14	75.87	
8	55.44	14.32	88.14

Table: Dissemilarity matrix for example data after step 6 using the average method.

Example

	1/5	2/4/6/8/10
2/4/6/8/10	46.24	
3/7/9	34.14	75.87

Table: Dissemilarity matrix for example data after step 7 using the average method.


Customer Data: Annual Income vs. Spending Score
Spending Score
75
50
25
0
15.0
17.5
20.0
22.5
Annual Income (k$)
Customer_ID
1
2
3
4
5
6
7
8
9
10

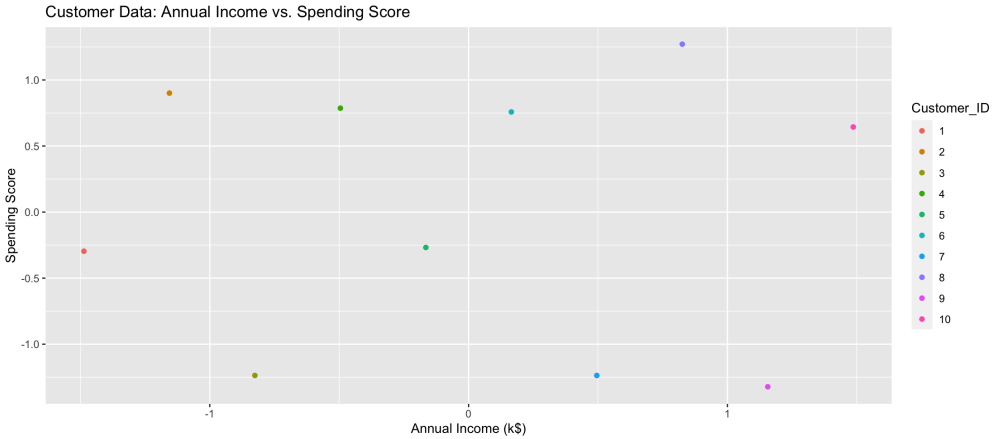
Example

	1	2	3	4	5	6	7	8	9
2	1.24								
3	1.15	2.16							
4	1.47	0.67	2.05						
5	1.32	1.53	1.17	1.10					
6	1.96	1.33	2.23	0.66	1.08				
7	2.19	2.70	1.32	2.25	1.17	2.02			
8	2.79	2.02	3.00	1.41	1.83	0.84	2.53		
9	2.83	3.21	1.98	2.68	1.69	2.30	0.67	2.61	
10	3.12	2.65	2.98	1.99	1.89	1.33	2.13	0.91	1.99

Table: Dissemilarity matrix for example data using Z-scores

Hierarchical Example

Customer ID	Annual Income (Z-score)	Spending Score (Z-score)
1	-1.49	-0.30
2	-1.16	0.90
3	-0.83	-1.24
4	-0.50	0.79
5	-0.17	-0.27
6	0.17	0.76
7	0.50	-1.24
8	0.83	1.27
9	1.16	-1.32
10	1.49	0.64


Customer Data: Annual Income vs. Spending Score
Spending Score
1.0
0.5
0.0
-0.5
-1.0
-1
0
1
Annual Income (k$)
Customer_ID
1
2
3
4
5
6
7
8
9
10

Example

	1	2	3	4/6	5	7	8	9
2	1.24							
3	1.15	2.16						
4/6	1.71	1.00	2.14					
5	1.32	1.53	1.17	1.09				
7	2.19	2.70	1.32	2.14	1.17			
8	2.79	2.02	3.00	1.12	1.83	2.53		
9	2.83	3.21	1.98	2.49	1.69	0.67	2.61	
10	3.12	2.65	2.98	1.66	1.89	2.13	0.91	1.99

Table: Dissemilarity matrix for example data after step 1 using the average method.

Example

	1	2	3	4/6	5	7/9	8
2	1.24						
3	1.15	2.16					
4/6	1.71	1.00	2.14				
5	1.32	1.53	1.17	1.09			
7/9	2.51	2.95	1.65	2.31	1.43		
8	2.79	2.02	3.00	1.12	1.83	2.57	
10	3.12	2.65	2.98	1.66	1.89	2.06	0.91

Table: Dissemilarity matrix for example data after step 2 using the average method.

Example

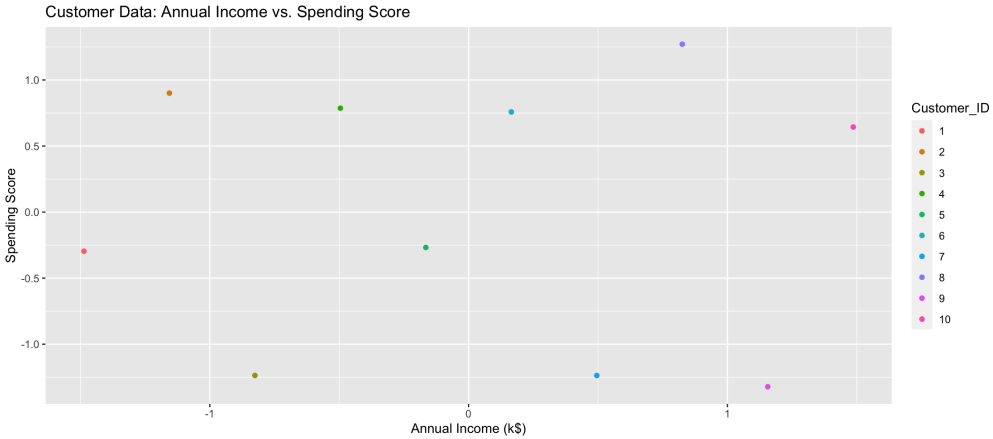
	1	2	3	4/6	5	7/9
2	1.24					
3	1.15	2.16				
4/6	1.71	1.00	2.14			
5	1.32	1.53	1.17	1.09		
7/9	2.51	2.95	1.65	2.31	1.43	
8/10	2.96	2.34	2.99	1.39	1.86	2.32

Table: Dissemilarity matrix for example data after step 3 using the average method.

Example

	1	2/4/6	3	5	7/9
2/4/6	1.48				
3	1.15	2.15			
5	1.32	1.31	1.17		
7/9	2.51	2.63	1.65	1.43	
8/10	2.96	1.86	2.99	1.86	2.32

Table: Dissemilarity matrix for example data after step 4 using the average method.


Customer Data: Annual Income vs. Spending Score
Spending Score
1.0
0.5
0.0
-0.5
-1.0
-1
0
1
Annual Income (k$)
Customer_ID
1
2
3
4
5
6
7
8
9
10

Example

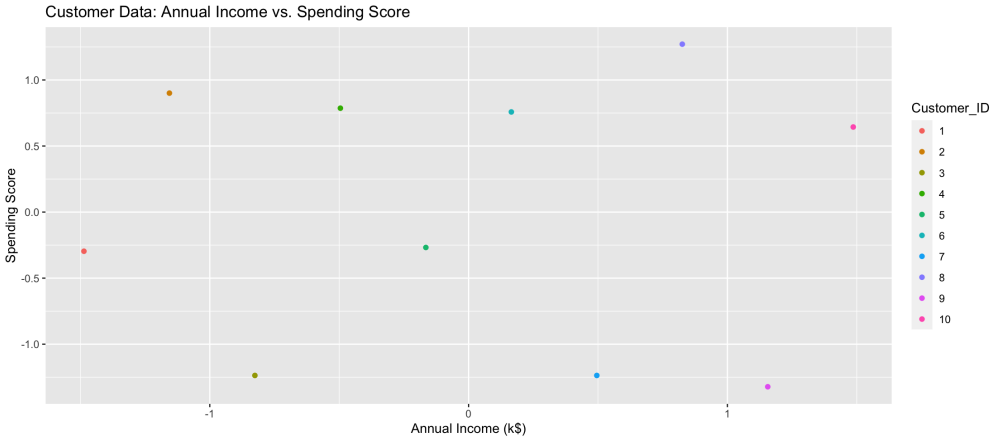
	1/3	2/4/6	5	7/9
2/4/6	1.81			
5	1.25	1.31		
7/9	2.08	2.63	1.43	
8/10	2.97	1.86	1.86	2.32

Table: Dissemilarity matrix for example data after step 5 using the average method.

Example

	1/3/5	2/4/6	7/9
2/4/6	1.56		
7/9	1.76	2.63	
8/10	2.42	1.86	2.32

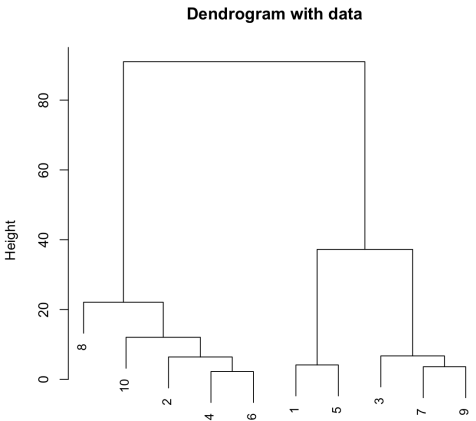
Table: Dissemilarity matrix for example data after step 6 using the average method.

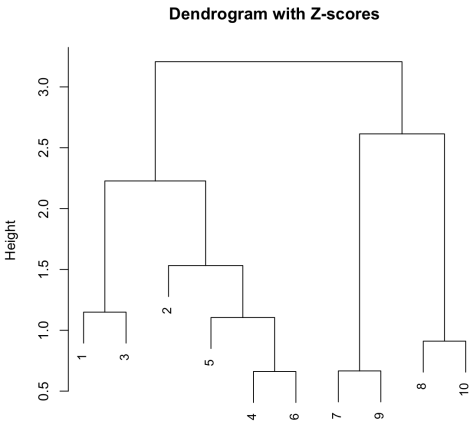

Customer Data: Annual Income vs. Spending Score
Spending Score
1.0
0.5
0.0
-0.5
-1.0
-1
0
1
Annual Income (k$)
Customer_ID
1
2
3
4
5
6
7
8
9
10

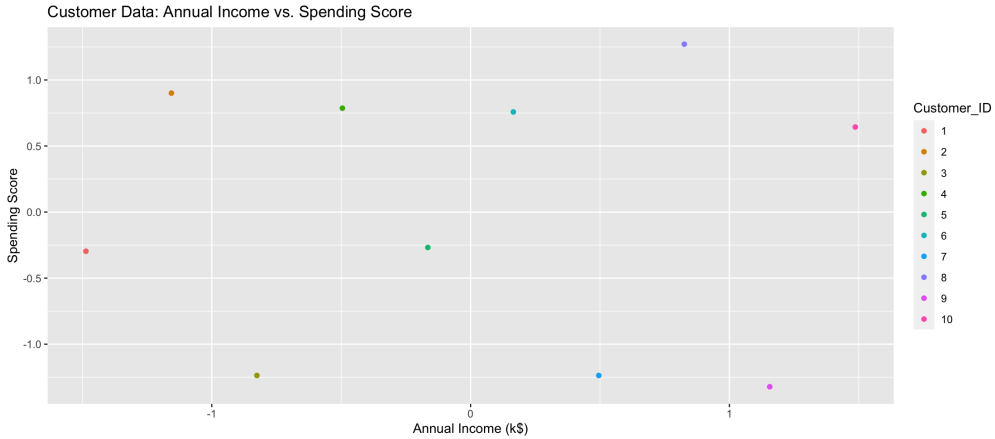
Example

	1/2/3/4/5/6	7/9
7/9	2.20	
8/10	2.14	2.32

Table: Dissemilarity matrix for example data after step 7 using the average method.


Dendrogram with data
Height
0
20
40
60
80
8
10
2
4
6
1
5
3
7
9


Dendrogram with Z-scores
Height
0.5
1.0
1.5
2.0
2.5
3.0
1
3
2
5
4
6
7
9
8
10


Customer Data: Annual Income vs. Spending Score
Spending Score
1.0
0.5
0.0
-0.5
-1.0
-1
0
1
Annual Income (k$)
Customer_ID
1
2
3
4
5
6
7
8
9
10

Principal components clustering

- Principal components clustering is simply clustering the data based on $M \leq p$ principal components.
- For example, we could cluster the food data based on the first two PCs