Topics:

1) Data reduction
   → reduce # of variables (dimensions of data) while preserving information
   - Principle Component Analysis
   - Multidimensional scaling

2) Cluster analysis
   → unsupervised - uses the structure of the data to create labels
   - Hierarchical clustering
   - k-means

3) Classification (Discriminant analysis)
   → supervised - uses existing labels to assign labels to new data
   - k-nearest neighbours
   - Linear and Quadratic discriminant analysis
   - Logistic regression

Multivariate data

→ m-dimensional data = m attributes (variables) for every object (data point)
   ↳ data point: $\tilde{x} = (x_1, \dots x_m)^T$

→ attributes
   - numerical continuous $x_i \in \mathbb{R}$
   - numerical discrete $x_i \in \mathbb{N}$
   - categorical x_i = gender

Ref: Multivariate data sets are represented by a matrix $X \in \mathbb{R}^{n \times m}$
   - n = # observations
   - m = # variables

$$X = \begin{bmatrix} \text{---} \ \tilde{x}_1^T \ \text{---} \\ \vdots \\ \text{---} \ x_m^T \ \text{---} \end{bmatrix} \quad \text{where} \quad \tilde{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{im} \end{pmatrix} = i^{th} \text{ observation}$$

Def: A random variable is a mapping $X: \Omega \to \mathbb{R}$ which assigns real numbers to possible outcomes.

Def: For a random variable X we define its

1. probability mass function $f_X(x) := P(X = x)$
2. cumulative distribution function $F_X(x) := P(X \leq x)$

Def: The random variable X is said to be

1. discrete $\equiv$ $\mathrm{Im}(X)$ is countable
2. continuous $\equiv$ $\exists f_X: \mathbb{R} \to \mathbb{R}$ s.t. $F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$

$\hookrightarrow$ f_X is called the probability density function of X

Def: The expected value of a r.v. X is

1. X discrete ... $\mu = \mathbb{E}[X] := \sum_{x \in \mathrm{Im}(X)} x \cdot f_X(x)$
2. X continuous ... $\mu = \mathbb{E}[X] := \int_{-\infty}^{\infty} x\, f_X(x)\,dx$

Theorem (PNS): The expected value of a function $g(X)$ is given by

1. $\mathbb{E}[g(X)] = \sum_x g(x) f_X(x)$
2. $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx$

Def: Consider random variables $X_1, X_2, \dots X_m$. Define

(standard deviation $\downarrow$)

1. variance $\mathrm{Var}[X_i] := \mathbb{E}[(X_i - \mathbb{E}X_i)^2] = \mathbb{E}[X_i^2] - (\mathbb{E}X_i)^2 = \sigma^2$
2. covariance $\mathrm{Cov}[X_i, X_j] := \mathbb{E}[(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)] = \mathbb{E}[X_1 X_2] - \mathbb{E}X_1\,\mathbb{E}X_2$
3. correlation $\mathrm{Cor}[X_i, X_j] := \frac{\mathrm{Cov}[X_i, X_j]}{\sqrt{\mathrm{Var}[X_i]\cdot\mathrm{Var}[X_j]}} = \frac{\mathrm{Cov}[X_i, X_j]}{\sigma_i \cdot \sigma_j}$ ... normalized covariance
4. covariance matrix of $\tilde{X} = (X_1, \dots X_m)$

$\Sigma = \mathrm{Cov}[\tilde{X}]$ where $\Sigma_{ij} = \mathrm{Cov}[X_i, X_j]$ ... $\Sigma_{ii} = \mathrm{Var}[X_i]$

5. correlation matrix of $\tilde{X} = (X_1, \dots X_m)$

$C = \mathrm{Cor}[\tilde{X}]$ where $C_{ij} = \mathrm{Cor}[X_i, X_j]$ ... $C_{ii} = 1$

Fact: Correlation is scaled s.t. $-1 \leq \mathrm{Cor}[X_i, X_j] \leq 1$

Def: X, Y are independent, $X \perp Y \equiv P[X=x \ \& \ Y=y] = P[X=x] \cdot P[Y=y]$

Theorem: If X and Y are independent, then

$$\mathbb{E}XY = \mathbb{E}X\,\mathbb{E}Y \quad \Rightarrow \quad \mathrm{Cov}(X,Y) = \mathrm{Corr}(X,Y) = 0$$

Theorem: Linear combinations of random variables give

1. $\mathbb{E}[aX+b] = a\mathbb{E}X + b$
2. $\mathbb{E}[aX+bY] = a\mathbb{E}X + b\mathbb{E}Y$
3. $\mathrm{Var}[aX+b] = a^2 \mathrm{Var}X$
4. $\mathrm{Var}[aX+bY] = a^2\mathrm{Var}X + b^2\mathrm{Var}Y + \underbrace{2ab\,\mathrm{Cov}[X,Y]}_{0 \text{ if } X \perp Y}$
5. $\mathrm{Cov}[aX+b, cY+d] = ac \cdot \mathrm{Cov}[X,Y]$
6. $\mathrm{Cov}[aX+bW, cY+dZ] = ac\,\mathrm{Cov}[X,Y] + ad \cdot \mathrm{Cov}[X,Z] + bc \cdot \mathrm{Cov}[W,Y] + bd \cdot \mathrm{Cov}[W,Z]$

Matrix representation

- **Expected value**

$$\mathbb{E}[a_1X_1 + \dots a_mX_m] = a_1\mu_1 + \dots + a_m\mu_m$$

$$a = (a_1, \dots a_m)^T \in \mathbb{R}^m$$
$$\tilde{X} = (X_1, \dots X_m)^T$$
$$\tilde{\mu} = (\mu_1, \dots \mu_m)^T$$

$$\Rightarrow \underline{\mathbb{E}[a^T\tilde{X}] = a^T\tilde{\mu}}$$

- **Variance**

$$\mathrm{Var}[a_1X_1 + \dots + a_mX_m] = \sum_{i=1}^{m} a_i^2 \underbrace{\mathrm{Var}X_i}_{\Sigma_{ii}} + \sum_{\substack{i,j=1 \\ i \neq j}}^{m} a_i a_j \underbrace{\mathrm{Cov}[X_i,X_j]}_{\Sigma_{ij}}$$

$$\Rightarrow \underline{\mathrm{Var}[a^T\tilde{X}] = a^T\Sigma a}$$

- **Covariance**

$$\mathrm{Cov}[a_1X_1 + \dots + a_mX_m, b_1X_1 + \dots + b_mX_m] = \sum_{i,j=1}^{m} a_i b_j \mathrm{Cov}[X_i,X_j]$$

$$\Rightarrow \underline{\mathrm{Cov}[a^T\tilde{X}, b^T\tilde{X}] = a^T\Sigma b = b^T\Sigma a}$$

Principle Component Analysis - PCA

Def: $\lambda \in \mathbb{R}$ is an eigenvalue of $A \in \mathbb{R}^{m \times m} \equiv \exists v \neq 0$ s.t. $Av = \lambda v$ → eigenvector

Finding eigenvalues: $Av = \lambda v \Rightarrow Av - \lambda v = 0 \Rightarrow (A - \lambda I)v = 0 \rightarrow \det(A - \lambda I) \neq 0$ (*)

(*) if $\det(A - \lambda I) = 0$, then $v = 0$ is the only solution (x)

Def: v is an unit eigenvector $\equiv \|v\| = 1$ ... standard norm $\|v\| = \sqrt{v^T v}$

Def: vectors $u, v \in \mathbb{R}^n$ are

- orthogonal $u \perp v \equiv u^T v = 0$
- orthonormal $\equiv u \perp v$ & $\|u\| = \|v\| = 1$

• Properties of the covariance matrix

Theorem: $\mathrm{Cov}[\tilde{X}] = \Sigma$ is symmetric and positive semi-definite.

Pf: Symmetric because $\mathrm{Cov}[X_i, X_j] = \mathrm{Cov}[X_j, X_i]$

note: $\mathrm{Var}[a^T \tilde{X}] = a^T \Sigma a$

↳ variance is always $\geq 0 \Rightarrow \Sigma$ is positive semi-definite ▨

Corollary: The eigenvalues of Σ are all non-negative

Pf: $\Sigma v = \lambda v \Rightarrow v^T \Sigma v = v^T \lambda v = \lambda v^T v \Rightarrow \lambda = \frac{v^T \Sigma v}{v^T v} \geq 0$

Theorem: $\Sigma \in \mathbb{R}^{m \times m}$ has m orthonormal eigenvectors.

Pf: Σ is symmetric $\Rightarrow \exists R \in \mathbb{R}^{m \times m}$ s.t. $R^T \Sigma R$ is diagonal & $R^T R = I_m$

↳ spectral decomp. of a symmetric matrix

- since Σ is diagonalizable, it has m lin. ind. eigenvectors
- they are orthonormal because $R^T R = I_m$ and the columns of R = eigenvectors of Σ ▨

• PCA

→ idea: our variables might not be conveying the information hidden in the data very efficiently

⇒ it might be possible to express most of the info with just a few carefully chosen linear combinations of the variables

→ if two variables are highly correlated, then we really only need one (kinda)

Goal: Describe the variation in a set of correlated variables $X_1, \dots X_m$ using a new set of uncorrelated variables $Y_1, \dots Y_p$ and hopefully $p \ll m$

Y_1 = lin. comb. of $\tilde{X}$ such that it accounts for the most variation possible

Y_2 = lin. comb. of $\tilde{X}$ accounting for as much of the remaining variance while subject to the constrain $\text{Corr}(Y_1, Y_2) = 0$ ... $Y_1 \perp Y_2$

Def: We have data: n observations of m variables $\tilde{X}_1, \tilde{X}_2, \dots \tilde{X}_m \in \mathbb{R}^m$

① sample mean of each variable $\overline{X_j} := \frac{1}{n} \sum_{i=1}^{n} X_{ij}$

② sample covariance matrix $Q \in \mathbb{R}^{m \times m}$

$$q_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (X_{ki} - \overline{X_i})(X_{kj} - \overline{X_j})$$

↳ k^{th} observation of variables X_i and X_j

Fact: Previously discussed results for Σ also apply for Q

Note: We divided by $n-1$ rather than n. This is because by using the sample mean, we have lost a degree of freedom. If we knew EX_i and EX_j, then we would divide by n.

→ consider a sample of X_i of size n. It exists in $\mathbb{R}^n$ and so has n degrees of freedom in movement - each sample member can be anything

→ however by fixing the sample mean, we are constraining the sample to have a fixed sum

⇒ $n-1$ members can be anything, but the last one must have the proper value to ensure the sample mean is unchanged

→ we are summing over some values n times, but it is in truth a sum of only $n-1$ independent things

linear combinations of data

→ let Y be lin comb of our m variables $Y = \sum_{i=1}^{m} a_i X_i = a^T \tilde{X}$

→ to determine $\mathrm{Var}\, Y$ we would need $\Sigma = \mathrm{Cov}(\tilde{X})$

⇒ estimate Σ with Q

⇒ $\underline{\mathrm{Var}(a^T \tilde{X}) \approx a^T Q a}$, $\underline{\mathrm{Cov}(a^T \tilde{X}, b^T \tilde{X}) \approx a^T Q b = b^T Q a}$

Problem: Find $a \in \mathbb{R}^m$ s.t. maximize $\mathrm{Var}(a^T \tilde{X}) \approx a^T Q a$ subject to $a^T a = 1$.

Solution: Use Lagrange multipliers. We want to

maximize $f : \mathbb{R}^m \to \mathbb{R}^+$, $\tilde{a} \mapsto a^T Q a$ subject to $g(\tilde{a}) := a^T a - 1 = 0$

Theorem: A necessary condition for $\tilde{a}$ being a local maximum of f subject to $g(\tilde{a}) = 0$ is existence of $\lambda \in \mathbb{R}$ s.t.

$$\nabla f(\tilde{a}) = \lambda \nabla g(\tilde{a})$$

⇒ define $L(\tilde{a}) := f(\tilde{a}) - \lambda g(\tilde{a}) = a^T Q a - \lambda (a^T a - 1)$ ... we want $\nabla L = 0$

$$L = \sum_{i,j=1}^{m} a_i a_j q_{ij} - \lambda \sum_{i=1}^{m} a_i^2 + 1$$

$$\frac{\partial L}{\partial a_k} = 2 a_k q_{kk} + \sum_{j \neq k} a_j q_{kj} + \sum_{i \neq k} a_i q_{ik} - 2\lambda a_k$$

$$= 2 a_k q_{kk} + 2 \sum_{j \neq k} a_j q_{kj} - 2\lambda a_k \quad \text{... } Q \text{ is symmetric}$$

$$\Rightarrow \frac{\partial L}{\partial a_k} = 0 \iff \sum_{j=1}^{m} a_j q_{kj} - \lambda a_k = 0$$

↳ $(q_{k1}, q_{k2}, \dots q_{km}) = k^{\text{th}}$ row of Q

$$\Rightarrow \nabla L = 0 \iff Q\tilde{a} = \lambda \tilde{a} \iff \lambda \text{ is an eigenvalue of } Q,\ \tilde{a} \text{ eigenvector}$$

Conclusion: $\tilde{a}$ <u>maximizes</u> $\mathrm{Var}(a^T \tilde{X})$ $\iff$ $\tilde{a}$ is an <u>eigenvector</u> of Q

↳ also since we require $a^T a = 0$, we want unit eigenvectors

$$\mathrm{Var}(a^T X) = a^T Q a = a^T \lambda a = \lambda a^T a = \lambda \Rightarrow \underline{\text{variance = eigenvalue}}$$

Consider $\tilde{a} \perp \tilde{b}$ eigenvectors of Q with eigenvalues λ_a, λ_b

$$\mathrm{Cov}(a^T \tilde{X}, b^T \tilde{X}) = a^T Q b = a^T \lambda_b b = \lambda_b a^T b = 0$$

↳ eigenvectors of Q are orthogonal

$\Rightarrow Y_1 = a^T \tilde{X}$ and $Y_2 = b^T \tilde{X}$ are <u>independent</u>

• **Principle components**

Method: Given n observations $\tilde{x}_1, \dots \tilde{x}_n \in \mathbb{R}^m$ of m variables $X_1, \dots X_m$

1) compute the covariance matrix of $\tilde{x}_1, \dots \tilde{x}_n \rightarrow Q \in \mathbb{R}^{m \times m}$

2) calculate the eigenvalues and eigenvectors of Q

$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$ $\qquad \forall i: a_i^T a_i = 1, \quad \forall i \neq j: \tilde{a}_i \perp \tilde{a}_j$

$\tilde{a}_1, \tilde{a}_2, \dots \tilde{a}_m$

3) the principle components of the data are variables

$Y_1 = a_1^T \tilde{X}, \quad Y_2 = a_2^T \tilde{X}, \quad \dots \quad Y_m = a_m^T \tilde{X}, \quad \forall i \neq j: Y_i \perp Y_j$

$\operatorname{Var} Y_1 = \lambda_1 \geq \operatorname{Var} Y_2 = \lambda_2 \geq \dots \geq \operatorname{Var} Y_m = \lambda_m$

4) transform the data matrix of $X_1, \dots X_m$ to a data matrix of $Y_1, \dots Y_m$

$\rightarrow$ suppose $Y_j = a_j^T \tilde{X}$ and observation $\tilde{x}_i \in \mathbb{R}^m \Rightarrow \tilde{y}_j = a_j^T \tilde{x}_i$

$$\mathbb{X} = \begin{pmatrix} — \ \tilde{x}_1^T \ — \\ — \ \tilde{x}_2^T \ — \\ \vdots \\ — \ \tilde{x}_n^T \ — \end{pmatrix} \in \mathbb{R}^{n \times m} \rightsquigarrow \mathbb{Y} = \mathbb{X} \cdot \begin{bmatrix} | & | & & | \\ \tilde{a}_1 & \tilde{a}_2 & \cdots & \tilde{a}_m \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{n \times m}$$

Theorem: $\operatorname{Var} X_1 + \dots + \operatorname{Var} X_m = \operatorname{Var} Y_1 + \dots + \operatorname{Var} Y_m = \sum_i \lambda_i$

Proof: We will show

$$\operatorname{trace}(Q) = \sum \operatorname{Var} X_i = \sum \operatorname{Var} Y_i = \sum \lambda_i$$

$\rightarrow$ consider the char. polynomial of Q

$$p_Q(t) = b_m t^m + b_{m-1} t^{m-1} + \dots + b_0 = \det \begin{bmatrix} q_{11} - t & q_{12} & \cdots & q_{1m} \\ q_{21} & q_{22} - t & & \vdots \\ \vdots & & \ddots & \\ q_{m1} & \cdots & & q_{mm} - t \end{bmatrix} = \prod_{i=1}^{m} (a_{ii} - t) + \dots$$

👁: $b_{m-1} = (-1)^{m-1} \sum_{i=1}^{m} a_{ii} = (-1)^{m-1} \operatorname{tr}(Q)$

$\rightarrow$ since there are m lin. ind. eigenvectors: $m = \sum \text{Geom}\, \lambda \leq \sum \text{Alg}\, \lambda \leq m$

$$p_Q(t) = (\lambda_1 - t)^{r_1} (\lambda_2 - t)^{r_2} \cdots (\lambda_s - t)^{r_s}, \qquad r_1 + r_2 + \dots + r_s = m = \sum \text{Alg}\, \lambda$$

👁 $b_{m-1} = (-1)^{m-1} \sum_{i=1}^{s} r_i \lambda_i$

$\Rightarrow$ when Q has eigenvalues (not necessarily distinct) $\lambda_1, \dots \lambda_m$: $\operatorname{tr} Q = \sum_i \lambda_i$ ∎

5) using this theorem we can see that

$$\frac{\lambda_k}{\sum_i \lambda_i} = \text{proportion of variance explained by } Y_k$$

• Scaling the data

Problem: The Iris dataset gives flower measurements in cm

→ would the results of PCA be different if one of the measurements was in mm^2?

→ yes, because PCA seeks to maximize variance and

$$Var(aX) = a^2 Var X \Rightarrow$$ it is very sensitive to data scaling

→ even worse, what if X_1 measured time and X_2 length?

solution: We want comparable units

⇒ make each variable have variance = 1

⇒ divide each variable by its std. dev: $Var\left(\frac{1}{\sigma} X\right) = \frac{1}{\sigma^2} Var X = 1$

⇒ do PCA on the Cov. matrix of the transformed data

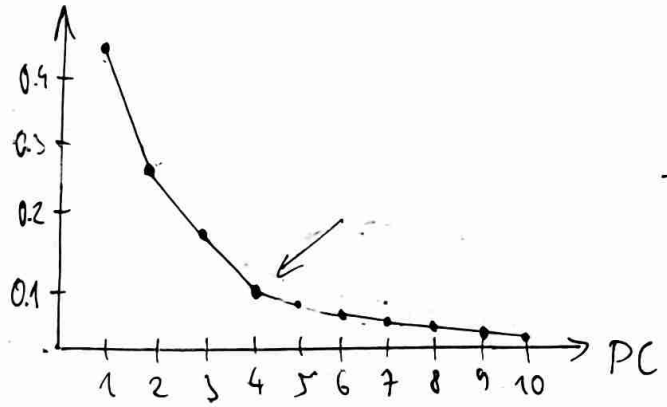
↳ this is the same as doing PCA with the Corr. matrix from the start

→ since $\forall i: Var X_i = 1$ we have $\sum_i \lambda_i = \sum_i Var X_i = m$

• How many PCs to choose?

1) keep adding until a fixed proportion of variance (90%) is included

2) look at the Scree plot



y = proportion of variance explained

→ look for an elbow

⇒ here we probably want to go with the first 4 components

When can PCA be used?

→ continuous variables

→ or numeric variables which could be interpreted as continuous

→ can't be applied for categorical variables (doesn't make sense)

Hierarchical Clustering

Idea: combine data to clusters based on how similar they are to each other
↪ start with every data point in its own cluster
→ gradually combine them
⇒ we need a dissimilarity measure between clusters

Def: The function $d(\tilde{x}, \tilde{y})$ is a metric $\equiv$

i) $d(\tilde{x}, \tilde{y}) \geq 0$ & $d(\tilde{x}, \tilde{y}) = 0 \iff \tilde{x} = \tilde{y}$

ii) $d(\tilde{x}, \tilde{y}) = d(\tilde{y}, \tilde{x})$

iii) $d(\tilde{x}, \tilde{y}) \leq d(\tilde{x}, \tilde{z}) + d(\tilde{z}, \tilde{y})$ ... in our case might be ignored

Examples:

- Euclidean $= \left(\sum_i (x_i - y_i)^2 \right)^{\frac{1}{2}}$
- Manhattan $= \sum_i |x_i - y_i|$
- Maximum $= \max_i |x_i - y_i|$
- Minkowski $= \left(\sum_i |x_i - y_i|^h \right)^{\frac{1}{h}}$, $h \geq 1$

Metrics for binary data

→ look at a cross tabulation of $\tilde{x}$ and $\tilde{y}$ and considering how much they agree / disagree

Point $\tilde{y}$

Point $\tilde{x}$	1	0	
1	a	b	a+b
0	c	d	c+d
	a+c	b+d	a+b+c+d

$x = (1,1,0,0,0,0,1)$
$y = (1,0,1,1,1,0,0)$

x \ y	1	0	
1	1	2	3 = #1 in x
0	3	1	4 = #0 in x
	4	3	7

- Hamming $= 1 - \frac{a+d}{a+b+c+d}$
- Jaccard $= 1 - \frac{a}{a+b+c}$ ... ignore double absence, as may be a redundant variable
- Kulczynski $= 1 - \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right)$ ... average of ratios of agreement from two samples
- Czekanowski $= 1 - \frac{2a}{2a+b+c}$ ... more emphasis on double presence

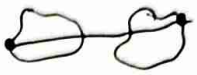
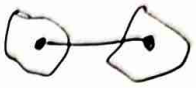
Categorical data

- we can simply def $d(\tilde{x}, \tilde{y}) := \#$ categories where $\tilde{x}$ and $\tilde{y}$ don't agree

Mixed data

- calculate dissimilarities for continuous, binary and categorical variables separately
- $\Rightarrow$ then do a weighted combination

Def: The dissimilarity of two groups $A = \{x_1, \dots x_k\}$ and $B = \{y_1, \dots y_l\}$ is a linkage

- Single linkage $= \min_{i,j} d(x_i, y_j)$  ... doesn't hold Δ-ineq
- Complete linkage $= \max_{i,j} d(x_i, y_j)$ 
- Average linkage $= \frac{1}{|A| \cdot |B|} \sum_{i,j} d(x_i, y_j)$ 

Example: Use manhattan metric and complete linkage to cluster the following data

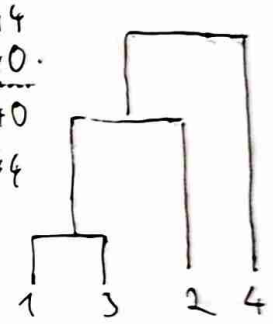
Observation	A	B	C	D
1	11	-6	-4	8
2	15	6	6	9
3	13	-5	-8	10
4	-12	5	-7	6

$\Rightarrow$ ← variables

dissimilarity matrix

	2	3	4
1	27	9	39
2		28	44
3			40
{1,3}	28		40
{1,2,3}			44

→ first we unify 1 and 3 to {1,3}
→ second 2 and {1,3} into {1,2,3}
→ lastly {1,2,3} and 4



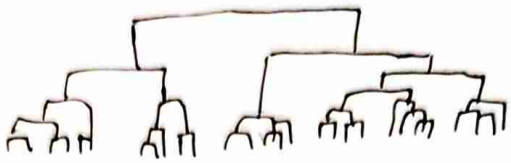
- Dendrogram
  - → the tree-like graph used to visualize hierarchical clustering is called a dendrogram
  - → joining clusters ~ U-link
    - ↪ lenght of the two legs of the U-link = distance between clusters

- Scaling the data
  - → we should scale the data before constructing the dissimilarity matrix
  - → if the variables are not scaled, then the variable with the greatest variance will dominate the distances and figure most prominently in the clustering solution
  - $\Rightarrow$ we divide the variable by its std. dev. $\Rightarrow$ Var = 1 $\Rightarrow$ then scaling is not good

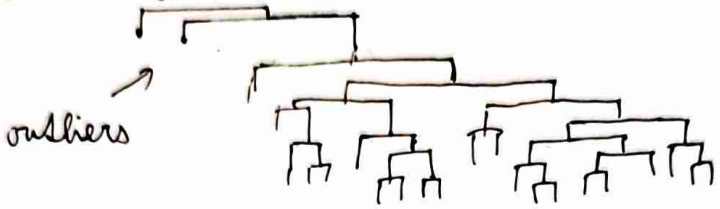
! sometimes we might want to give less weight to a variable which carries less info (has small variance)

Linkage effects

- complete linkage joins the final clusters at a larger measure of dissimilarity
- complete and average linkage result in "spherical" clusters with good internal similarity



- single linkage displays outliers, which are often hidden in complete linkage



- complete and single linkage are invariant under monotonic transformations of the dissimilarity matrix entries, while average linkage is not
- complete linkage is likely to suggest a smaller number of large clusters

! we should first give careful thought to which linkage makes sense
↳ trying them all and deciding based on which one looks the best can easily turn an objective solution to a subjective one

Chaining

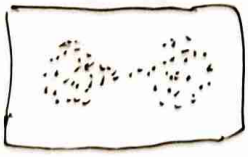
- phenomenon which occurs while using single linkage

→ single linkage has the tendency to repeatedly add a single observation to the same group that continues to get larger and larger

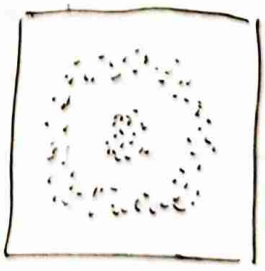
→ occurs because a unit joins a group based on similarity with just one member of that group

→ this results in elongated clusters that may include quite dissimilar points

→ however, chaining is not always bad



- average works ✓
- single chains all together



- average/complete tries to create spherical clusters and fails
- single manages to make 2 clear clusters thanks to chaining

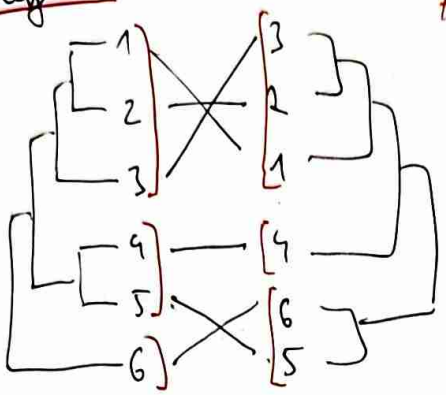
Cluster Agreement

→ two different clustering methods were applied to the same data

- single × complete linkage hierarchical clustering
- expert uses his knowledge × statistician uses data

→ can we quantify the level of agreement between the two approaches?

- **Tanglegrams**



clusters = 3

→ the two methods agree on the first cluster

→ but they disagree on the other two

! they might disagree with the same # clusters but agree on different

- **Cross tabulation**

Method A \ Method B	C_1	C_2	
C_1	80	10	90
C_2	15	60	75
	95	70	115

agree on $\frac{90}{115}$

	C_1	C_2	
C_1	10	30	40
C_2	60	15	75
	70	45	115

same as the first table → just rename C_1 and C_2

	C_1	C_2	
C_1	24	16	40
C_2	46	29	75
	70	45	115

→ don't really agree on anything

$\frac{53}{115}$ and $\frac{62}{115}$ are both meh

	C_1	C_2	
C_1	10	30	40
C_2	20	5	25
C_3	40	10	50
	70	45	115

→

	C_1	C_2	
C_1	10	30	40
C_2+C_3	60	15	75
	70	45	115

→ this is in fact the same table as in example 2

→ in more complex cases it is very difficult to decide agreement just by looking

⇒ in 1971 has Rand proposed an index for measuring the agreement as a number between 0 and 1

• The Rand Index

A \ B	C_1	C_2	$\cdots$	C_ℓ	
C_1	n_{11}	n_{12}	$\cdots$	$n_{1\ell}$	n_{1*}
C_2	n_{21}	n_{22}	$\cdots$	$n_{2\ell}$	n_{2*}
$\vdots$	$\vdots$	$\vdots$	$\ddots$		$\vdots$
C_k	n_{k1}	n_{k2}		$n_{k\ell}$	n_{k*}
	n_{*1}	n_{*2}	$\cdots$	$n_{*\ell}$	n

→ n observations $\tilde{x}_1, \dots \tilde{x}_n$

$\alpha := \#(\tilde{x}_i, \tilde{x}_j)$: method A: same cluster, method B: same cluster

$\beta := \#(\tilde{x}_i, \tilde{x}_j)$: method A: different clusters, method B: different clusters

$\gamma := \#(\tilde{x}_i, \tilde{x}_j)$: A: same, B: different

$\delta := \#(\tilde{x}_i, \tilde{x}_j)$: A: different, B: same

👁 # all possible pairs of data points $\binom{n}{2} = \alpha + \beta + \gamma + \delta$

$$\Rightarrow \text{Rand index } R = \frac{\alpha+\beta}{\alpha+\beta+\gamma+\delta} = \frac{\binom{n}{2} - \gamma - \delta}{\binom{n}{2}}$$

- $\gamma = \sum_{i=1}^{k} \binom{n_{i*}}{2} - \sum_{i,j} \binom{n_{ij}}{2}$
- $\delta = \sum_{j=1}^{\ell} \binom{n_{*j}}{2} - \sum_{i,j} \binom{n_{ij}}{2}$

- same / diff = same / * − same / same
- diff / same = * / same − same / same

→ the rand index of the tables on previous page is

1: 0.654 2: 0.654 3: 0.499 4: 0.588

❗ the value for table 3 seems quite large, but there wasn't much agreement

Example: Calculate the rand index of

	C_1	C_2	
C_1	5	25	30
C_2	20	5	25
C_3	40	5	45
	65	35	100

$$\alpha = \sum_{i,j} \binom{n_{ij}}{2} = 3\cdot\binom{5}{2} + \binom{25}{2} + \binom{20}{2} + \binom{40}{2} = 1300$$

$$\gamma = \sum_{i} \binom{n_{i*}}{2} - \alpha = \binom{30}{2} + \binom{25}{2} + \binom{45}{2} - 1300 = 425$$

$$\delta = \sum_{j} \binom{n_{*j}}{2} - \alpha = \binom{65}{2} + \binom{35}{2} - 1300 = 1375$$

$$\binom{n}{2} = \binom{100}{2} = 4950 \Rightarrow R = \frac{4950 - 425 - 1375}{4950} \approx 0.636$$

Problem: The Rand index tends to give large values even when the clustering methods are in substantial disagreement

• Adjusted Rand index

$$\alpha = \sum_{i,j} \binom{n_{ij}}{2}$$

$$\beta_A = \sum_{i} \binom{n_{i*}}{2}$$

$$\beta_B = \sum_{j} \binom{n_{*j}}{2}$$

$$ARI = \frac{\alpha\binom{n}{2} - \beta_A\beta_B}{\frac{1}{2}(\beta_A+\beta_B)\cdot\binom{n}{2} - \beta_A\beta_B} = \frac{RI - \mathbb{E}[RI]}{\text{Max } RI - \mathbb{E}[RI]}$$

↳ uses hypergeometric distribution

→ ARI can be negative, but it can't be greater than 1

→ for tables 1 to 4: 1: 0.311, 2: 0.311, 3: -0.006, 4: 0.185

↳ now we have a small value for table 3

→ ARI ≤ 0 ⇒ no agreement

• Classification with kNN = k-nearest neighbours

- we have data with labels and want to classify new data points
- kNN is non-parametric and doesn't make any assumptions on the spread of the data ↳ the distribution of the data

⇒ there is no measurement of uncertainty when assigning labels

- kNN simply looks at the k closest points and assigns the new point to the group which has the majority

- two things to consider

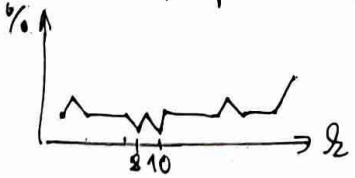
1) do we scale the data? results may vary
2) how will we calculate the distance?
3) how big should k be? ← important

- choosing k

→ split the data into

- training set – will be used to classify "unlabeled" data
- test set – treated as unlabeled and is used to find the best k
- validation set – treated as unlabeled, used to estimate the classification error of the best k

→ plot the proportion of incorrectly classified data from the test set



→ k = 8 has the best rate (8 < 10)

⇒ now test the correct classification rate on the validation set to estimate the error

→ this is a general technique

- training data – data used to fit several models
- test data – data used to compare these models and choose the best one
- validation data – used to access the performance of the chosen model

→ typical split is 50% 25% 25%

- **Cross-Validation**
  - another general performance assessment technique we can use to pick k
  - → do this for every model:
    1) split the data into k subsets $X_1, \dots X_k$
    2) For $i = 1 \dots k$:
       - fit the model using $\{X_1, \dots X_k\} \setminus \{X_i\}$
       - then count how many points from X_i would be correctly classified
    3) Calculate the total classification rate
  - ⇒ now pick the model with the best classification rate
- **leave-one-out cross-validation** = # subsets = # datapoints
- **k-fold cross-validation** = # subsets = k

Discriminant Analysis - classification

- supervised statistical techniques where we assume some info about the classes and use that to classify new data
- used when we know that there are k groups within the data and that there is a subset of the data which is labeled
- Linear DA (LDA) and Quadratic DA (QDA) both assume a distribution over the data
  - ⇒ now we can use probability theory to calculate the probability of a point belonging to a group under the assumptions we have made

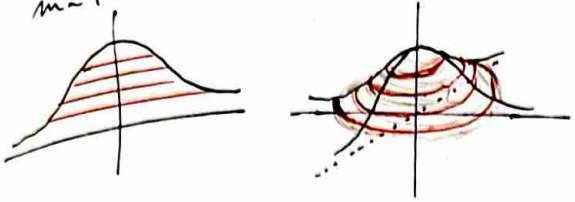
Def: Let $\tilde{X} = (X_1, \dots X_m)^T$ be a vector of random variables.
We say that $\tilde{X}$ follows a Multi-Variate Normal (MVN) distribution $\tilde{X} \sim MVN(\tilde{\mu}, \Sigma)$, where $\tilde{\mu} \in \mathbb{R}^m$ and $\Sigma \in \mathbb{R}^{m \times m}$ is positive semi-definite
$\equiv$ the probability-density function of $\tilde{X}$ is

$$f(\tilde{x} | \tilde{\mu}, \Sigma) = \left((2\pi)^m \cdot |\Sigma|\right)^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2}(\tilde{x} - \tilde{\mu})^T \Sigma^{-1} (\tilde{x} - \tilde{\mu})\right)$$

($|\Sigma| \hookrightarrow \det(\Sigma)$)

Theorem: The covariance matrix of $\tilde{X}$ is $\mathrm{Cov}(\tilde{X}) = \Sigma$.
The means of $\tilde{X}$ are given by $\tilde{\mu}$ ... $\mathbb{E}X_i = \mu_i$

Intuition:



→ the pdf has a bell shape

→ $\{\tilde{x} \mid f(\tilde{x}) = C\}$ makes circles / ellipses

→ $\{\tilde{x} \mid f(\tilde{x}) \geq C\}$ makes filled in ellipsoids

→ in general:

$$f(\tilde{x}) \geq C \iff \frac{1}{\left((2\pi)^m |\Sigma|\right)^{\frac{1}{2}}} \cdot \exp\left(-\frac{1}{2}(\tilde{x}-\tilde{\mu})^T \Sigma^{-1}(\tilde{x}-\tilde{\mu})\right) \geq C$$

$$\iff \exp(\cdots) \geq C(\ldots)^{\frac{1}{2}}$$

$$\iff -\frac{1}{2}(\tilde{x}-\tilde{\mu})^T \Sigma^{-1}(\tilde{x}-\tilde{\mu}) \geq \ln(\ldots)$$

$$\iff (\tilde{x}-\tilde{\mu})^T \Sigma^{-1}(\tilde{x}-\tilde{\mu}) \leq -2\ln\left(C\sqrt{(2\pi)^m |\Sigma|}\right)$$

$$\Rightarrow \{\tilde{x} \mid f(\tilde{x}) < C\} = \{\tilde{x} \mid (\tilde{x}-\tilde{\mu})^T \Sigma^{-1}(\tilde{x}-\tilde{\mu}) \leq \text{some number}\}$$

↳ which for positive-semidefinite Σ is an ellipsoid centered at $\tilde{\mu}$ }

→ if we assume that the data within group k follows $MVN(\tilde{\mu}_k, \Sigma_k)$, then the scatter of the data should be roughly elliptical

↳ $\tilde{\mu}_k$ ~ location of the ellipsoid & Σ_k ~ shape of the ellipsoid

• Distance approach

Def: The Mahalanobis distance of $\tilde{x} \in \mathbb{R}^m$ from the center $\tilde{\mu}$ is D, where

$$D^2 = (\tilde{x}-\tilde{\mu})^T \Sigma^{-1}(\tilde{x}-\tilde{\mu})$$

Two points $\tilde{x}_1$ and $\tilde{x}_2$ are on the same ellipsoid shell $\iff D_1 = D_2$

→ we can use the Mahalanobis distance to which cluster should $\tilde{x}$ belong

↳ want to find cluster k s.t. $(\tilde{x}-\tilde{\mu}_k)^T \Sigma_k^{-1}(\tilde{x}-\tilde{\mu}_k)$ is minimized

$\Rightarrow \tilde{x}$ is closer to cluster 1 than to cluster 2 $\iff$

$$(\tilde{x}-\tilde{\mu}_1)^T \Sigma_1^{-1}(\tilde{x}-\tilde{\mu}_1) < (\tilde{x}-\tilde{\mu}_2)^T \Sigma_2^{-1}(\tilde{x}-\tilde{\mu}_2)$$ ← Quadratic expression in $\tilde{x}$

→ if Σ is the same for all clusters, this simplifies: (Σ is symmetric $\Rightarrow \Sigma^{-1}$ is as well)

$$(\tilde{x}-\tilde{\mu}_1)^T \Sigma^{-1}(\tilde{x}-\tilde{\mu}_1) < (\tilde{x}-\tilde{\mu}_2)^T \Sigma^{-1}(\tilde{x}-\tilde{\mu}_2)$$

$$x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_1 - \mu_1^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}\mu_1 < x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_2 - \mu_2^T\Sigma^{-1}x + \mu_2^T\Sigma^{-1}\mu_2$$

$$-2x^T\Sigma^{-1}\mu_1 + \mu_1^T\Sigma^{-1}\mu_1 < -2x^T\Sigma^{-1}\mu_2 + \mu_2^T\Sigma^{-1}\mu_2$$

$$\Rightarrow x^T\Sigma^{-1}(\mu_1-\mu_2) > \frac{1}{2}\left(\mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2\right)$$ ← Linear expression in $\tilde{x}$

• Probabilistic approach

→ we want to classify a new observation $\tilde{x}$ into one of the K clusters

⇒ let $X: \mathbb{R}^m \to [K]$ be a random variable assigning points to clusters

Define $\pi_k := P[X=k]$ = proportion of population objects belonging to cluster k

→ Bayes theorem states

$$P[X=k|\tilde{x}] = \frac{P[X=k]\cdot P[\tilde{x}|X=k]}{P[\tilde{x}]} \quad \rightsquigarrow \quad P[X=k|\tilde{x}] \propto \pi_k P[\tilde{x}|X=k]$$

→ we are however dealing with continuous r.v., so we need to use the pdf:

$$P[X=k|\tilde{x}] \propto \pi_k \cdot f(\tilde{x}|X=k) = \pi_k \cdot f_k(\tilde{x})$$

→ we know the pdf for cluster k is from $MVN(\tilde{\mu}_k, \Sigma_k)$

→ we can calculate these probabilities by estimating $\tilde{\mu}_k$ and Σ_k

→ then we will assign $\tilde{x}$ to the cluster with the largest probability

$$P[X=k|\tilde{x}] > P[X=l|\tilde{x}] \iff \pi_k \cdot f_k(\tilde{x}) > \pi_l f_l(\tilde{x})$$

$$\iff \ln\pi_k + \ln f_k(\tilde{x}) > \ln\pi_l + \ln f_l(\tilde{x})$$

Recall: $f_k(\tilde{x}) = \left((2\pi)^m |\Sigma_k|\right)^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2}(\tilde{x}-\tilde{\mu}_k)^T \Sigma_k^{-1}(\tilde{x}-\tilde{\mu}_k)\right)$

$$\iff \ln\pi_k - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) > \ln\pi_l - \frac{1}{2}\ln|\Sigma_l| - \frac{1}{2}(x-\mu_l)^T\Sigma_l^{-1}(x-\mu_l)$$

① **Linear DA**: assumes all Σ for all clusters

$$\ln\pi_k - \frac{1}{2}\left[x^T\Sigma^{-1}x - 2x^T\Sigma^{-1}\mu_k + \mu_k^T\Sigma^{-1}\mu_k\right] > \ln\pi_l - \frac{1}{2}\left[x^T\Sigma^{-1}x - 2x^T\Sigma^{-1}\mu_l + \mu_l^T\Sigma^{-1}\mu_l\right]$$

$$\ln\pi_k + x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k > \ln\pi_l + x^T\Sigma^{-1}\mu_l - \frac{1}{2}\mu_l^T\Sigma\mu_l$$

$$\ln\frac{\pi_k}{\pi_l} + x^T\Sigma^{-1}(\mu_k-\mu_l) > \frac{1}{2}\left(\mu_k^T\Sigma^{-1}\mu_k - \mu_l^T\Sigma^{-1}\mu_l\right) \iff P[\tilde{x}\in k] > P[\tilde{x}\in l]$$

↳ if we assume $\forall k$: $\pi_k = \frac{1}{K}$, then $\ln\frac{\pi_k}{\pi_l} = 0$

and we get the formula we have gotten from the distance approach

② **Quadratic DA**: different clusters have different Σ_k

→ no simplification arises, so $P[\tilde{x}\in k] > P[\tilde{x}\in l] \iff$

$$\ln\pi_k - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(\tilde{x}-\tilde{\mu}_k)^T\Sigma_k^{-1}(\tilde{x}-\tilde{\mu}_k) > \ln\pi_l - \frac{1}{2}\ln|\Sigma_l| - \frac{1}{2}(\tilde{x}-\tilde{\mu}_l)^T\Sigma_l^{-1}(\tilde{x}-\tilde{\mu}_l)$$

• Estimating Covariances

→ labeled data classified into K groups of sizes $n_1, n_2, \dots n_K$

1. For $\forall$ group k: calculate its sample covariance matrix $Q_k \in \mathbb{R}^{m \times m}$ and its sample mean $\bar{x}_k \in \mathbb{R}^m$

- QDA: we are finished
- LDA: we assume that $\forall k,l: \Sigma_k = \Sigma_l$ ⇒ need to pool the matrices

→ Q_k has $(n_k - 1)$ degrees of freedom ... fixes $\tilde{\mu}_k$

→ the pooled matrix Q will have $N - K$ degrees of freedom ... $\mu_1, \dots \mu_K$

$$\Rightarrow Q = \frac{1}{N-K} \cdot \sum_{k=1}^{K} (n_k - 1) \cdot Q_k$$

(N data points in total)

2. Perform classification

→ define $\forall k$: (prior probabilities)

- $\pi_k := \frac{1}{K}$ or $\frac{n_k}{N}$ or something else
- $f_k(\tilde{x}) := \left((2\pi)^m |Q_k|\right)^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2}(\tilde{x} - \bar{x}_k)^T Q_k^{-1} (\tilde{x} - \bar{x}_k)\right)$

→ the estimated probability that $\tilde{x} \in$ group k is proportional to

$$P[k|\tilde{x}] \propto \pi_k \cdot f_k(\tilde{x})$$ ← posterior probabilities

→ assign $\tilde{x}$ to the group with the largest probability

→ the decision boundary between class k and class l is given by

$$\frac{P[k|\tilde{x}]}{P[l|\tilde{x}]} = \frac{\pi_k f_k(\tilde{x})}{\pi_l f_l(\tilde{x})} = 1 \quad \Rightarrow \quad \log \frac{P[k|\tilde{x}]}{P[l|\tilde{x}]} = \log \frac{\pi_k}{\pi_l} + \log \frac{f_k(\tilde{x})}{f_l(\tilde{x})} = 0$$

Summary:

- LDA and QDA are model based parametric classifiers where the data of each group is assumed to follow a MVN distribution
  - model based ⇒ we can estimate the probability of correct assignment
  - MVN ⇒ the groups are assumed to have an elliptical shape

LDA: all groups have the same covariance matrix

QDA: different covariance matrices between groups

- **k-means Clustering**

- simple algorithm for dividing data into k groups

1. initialize the cluster means $\tilde{\mu}_1, \dots \tilde{\mu}_k$
2. while not happy:
3. assign every point to the nearest cluster mean
4. recalculate the cluster means based on those assignments

→ how can we tell that it converging?

$SS := \sum (\tilde{x}_i - \mu(x_i))^2$ = sum of squared distances of each point to its centroid

⇒ keep iterating until

1) the assignments stop changing
2) the improvement in $SS \approx 0$

→ picking initial cluster centroids

→ the algorithm might converge to a local minima

1. initialize randomly and do multiple runs
2. select them based on prior knowledge of the data
3. perform hierarchical clustering first as a basis

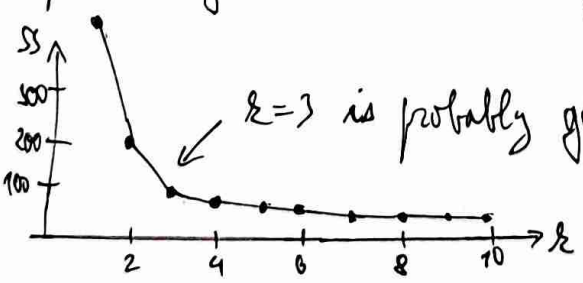
→ Choosing k

- what if we don't know how many clusters there should be?

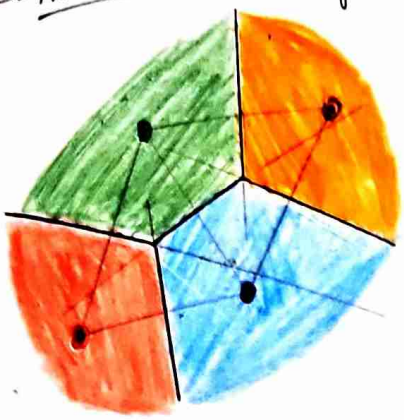
→ run for $k = 1, 2, \dots 10, \dots$ and try to minimize SS

! SS will naturaly decrease as k will go up

⇒ plot k against SS and look for an elbow in the graph



→ when k-means fails



- k-means classifies data by dividing the plane by lines

  ⇒ it cannot deal with data that don't have compact spherical groups

- Silhouette width

→ a general technique for evaluating the performance of a clustering solution

→ for each observation $\tilde{x}_i$ from $\tilde{x}_1, \dots \tilde{x}_n$ compute

- a_i := average distance from $\tilde{x}_i$ to the other points in its cluster
- b_i := average distance from $\tilde{x}_i$ to the nearest cluster it is not in

  ↳ calculate the dist for every cluster and take the minimum

- $s_i := \frac{b_i - a_i}{\max(a_i, b_i)}$ ← silhouette width of $\tilde{x}_i$

$-1 \leq s_i \leq 1$

- $s_i \simeq 1 \Rightarrow b_i \gg a_i \Rightarrow$ good cluster separation
- $s_i \simeq 0 \Rightarrow b_i \simeq a_i \Rightarrow$ clusters are poorly separated / overlapping
- $s_i < 0 \Rightarrow \tilde{x}_i$ has probably been assigned to the wrong cluster

→ we can look at the average silhouette width in each cluster and for the data overall

• Multidimensional Scaling - MDS

→ we have data $\tilde{x}_1, \dots \tilde{x}_m \in \mathbb{R}^m$

↳ distances $d_{ij} := d(\tilde{x}_i, \tilde{x}_j)$

⇒ we want to find $\tilde{y}_1, \dots \tilde{y}_m \in \mathbb{R}^d$, $d < m$

↳ distances $\delta_{ij} := d(\tilde{y}_i, \tilde{y}_j)$

such that δ_{ij} are as close to d_{ij} as possible for all i, j

⇒ object of MDS = provide an optimal configuration of observations in $\mathbb{R}^d$

→ we generally want a mapping $\delta_{ij} = f(d_{ij})$

- metric MDS: f continuous and monotonic
- non-metric MDS: f monotonic

$$d_{ij} < d_{kl} \Rightarrow \delta_{ij} < \delta_{kl}$$

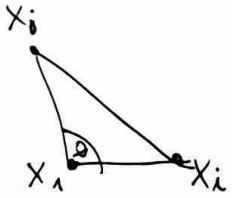
① Classical Metric Scaling

→ assume that f is the identity function

→ lets just say that we want to find an equivalent representation in $\mathbb{R}^m$ which is centered in the origin

→ we have calculated the distances d_{ij} and want to reconstruct $\tilde{x}_1 \dots \tilde{x}_m$

→ let $\tilde{x}_1$ be at the origin and consider $\tilde{x}_i$ and $\tilde{x}_j$



$$d_{ij}^2 = d_{1i}^2 + d_{1j}^2 - 2 d_{1i} d_{1j} \cos(\theta) \Rightarrow -\frac{1}{2}\left(d_{ij}^2 - d_{1i}^2 - d_{1j}^2\right) = d_{1i} d_{1j} \cos\theta$$

$$x_i^T x_j = \|x_i\| \cdot \|x_j\| \cdot \cos\theta = d_{1i} d_{1j} \cos\theta \quad , \text{ where } \|\cdot\| = \text{Euclid norm}$$

$$\Rightarrow \underbrace{x_i^T x_j}_{\text{want}} = \underbrace{-\frac{1}{2}\left(d_{ij}^2 - d_{1i}^2 - d_{1j}^2\right)}_{\text{know}}$$

⇒ we construct $B \in \mathbb{R}^{m \times m}$ s.t. $b_{ij} = -\frac{1}{2}\left(d_{ij}^2 - d_{1i}^2 - d_{1j}^2\right) = x_i^T x_j$

⇒ $B = X^T X$ for some X → k^{th} column of X is $\tilde{x}_k$

⇒ B is symmetric ⇒ as in PCA can be decomposed as

$$B = R D R^T = R\sqrt{D}\sqrt{D} R^T = R\sqrt{D}\sqrt{D}^T R^T = (R\sqrt{D}) \cdot (R\sqrt{D})^T$$

⇒ hence $X^T = R\sqrt{D}$, where R contains unit eigenvectors of B
and $\sqrt{D}$ contains the square roots of the eigenvalues

⇒ the vector $\tilde{x}_i$ can be expressed as

$$X^T = R\sqrt{D} \Rightarrow \begin{pmatrix} — & x_1^T & — \\ — & x_2^T & — \\ & \vdots & \\ — & x_m^T & — \end{pmatrix} = \begin{pmatrix} | & | & & | \\ \sqrt{\lambda_1}v_1 & \sqrt{\lambda_2}v_2 & \dots & \sqrt{\lambda_m}v_m \\ | & | & & | \end{pmatrix}$$

- $v_1, \dots v_m$ ↳ unit eigenvectors
- $\lambda_1, \dots \lambda_m$ ↳ eigenvalues

$$\Rightarrow \underline{\tilde{x}_{ji} = \sqrt{\lambda_i}\, v_{ij}}$$

→ the squared euclidean distance between $\tilde{x}_i$ and $\tilde{x}_j$ is

$$d_{ji}^2 = \sum_{i=1}^{m} (\tilde{x}_{ji} - \tilde{x}_{ki})^2 = \sum_{i=1}^{m} (\sqrt{\lambda_i}\, v_{ij} - \sqrt{\lambda_i}\, v_{ik})^2 = \underline{\sum_{i=1}^{m} \lambda_i (v_{ij} - v_{ik})}$$

⇒ lets say we want to shrink the data to $\tilde{y}_1, \dots \tilde{y}_m \in \mathbb{R}^d$, $d < m$

1. order the eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$
2. define $\tilde{y}_i \in \mathbb{R}^d$ as $\underline{\tilde{y}_{ji} := \sqrt{\lambda_i}\, v_{ij}}$
3. then the distance between $\tilde{y}_i$ and $\tilde{y}_j$ is

$$\delta_{ji}^2 = \sum_{i=1}^{d} (\tilde{y}_{ji} - \tilde{y}_{ki})^2 = \underline{\sum_{i=1}^{d} \lambda_i (v_{ij} - v_{ik})}$$

↳ using PCA knowledge

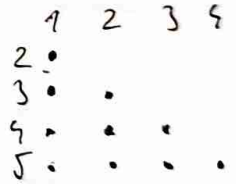
4. to choose appropriate d we can consider

$$\left(\sum_{i=1}^{d} \lambda_i\right) \Big/ \left(\sum_{i=1}^{m} \lambda_i\right) = \text{proportion of variance explained by } d \text{ dimensions}$$

- <u>performance evaluation</u>

→ we want the stress of the MDS to be as little as possible

$$\underline{\text{Stress}} := \sum_{i=2}^{m} \sum_{j<i} (d_{ij} - \delta_{ij})^2 \quad \dots \text{basically}$$



👁 it is better to consider the <u>relative error</u>

$\frac{(d_{ij} - \delta_{ij})^2}{d_{ij}}$ ⇒ small d_{ij} want better accuracy of δ_{ij} $\quad d=5,\ \delta=3$

! not the same → $d=50,\ \delta=48$

$$\underline{\text{Sammon Stress}} := \left(\sum_{i \neq j}^{} \frac{(d_{ij} - \delta_{ij})^2}{d_{ij}}\right) \Big/ \left(\sum_{i \neq j} d_{ij}\right)$$

(2) Metric least squares scaling

→ finds a configuration $\tilde{y}_1, \dots \tilde{y}_n \in \mathbb{R}^d$ which minimizes a loss function S e.g. stress or Samon-stress.

→ iterative numeric approach

👁 classical MDS uses the Enclidian distance model and minimizes the stress

(3) Non-metric multidimensional scaling

$$\text{Kruskal Stress} := \left(\sum_{i \neq j} \left(f(d_{ij}) - \delta_{ij}\right)^2\right) \Big/ \left(\sum_{i \neq j} \delta_{ij}^2\right)$$

→ it isn't trying achieve $\frac{d_{ij}}{\delta_{ij}} \approx 1$

→ it only seeks to preserve the rank order of the distances

$$d_{ij} < d_{kl} \Rightarrow \delta_{ij} < \delta_{kl} \qquad \forall i, j, k, l$$

→ iterative numeric algorithm again

→ good for non-metric ≈ non geometric data → ordinal data

→ to choose d we can plot the stress vs. d and look for an elbow

• Procrustes Analysis

👁 if we take a MDS configuration and rotate / translate / reflect it, all of the distances remain unchanged ⇒ it's an equivalent solution

⇒ Say two MDS methods have been applied to a set of n points $\in \mathbb{R}^m$ resulting in coordinate matrices $X \in \mathbb{R}^{n \times d}$ and $Y \in \mathbb{R}^{n \times d}$

↳ we want to know if they are equivalent

⇒ we want to match the i^{th} point in X to the i^{th} point in Y

⇒ minimize $R^2 := \sum_{i=1}^{n} (\tilde{y}_i - \tilde{x}_i)^T (\tilde{y}_i - \tilde{x}_i) = \sum_{i=1}^{n} \sum_{j=1}^{d} (y_{ij} - x_{ij})^2$

⇒ we will keep Y fixed (reference configuration) and transform X by rotating, translating and reflecting it to minimize R^2

→ we will also allow uniform scaling of the points in X - preserves distance ratios

→ the point $\tilde{x}_i$ will be transformed to

$$\tilde{x}_i \mapsto \tilde{x}_i' = s A^T \tilde{x}_i + b$$

where
- $s \in \mathbb{R}$ ... scaling factor note: $s \cdot I_d$ is called the dilation matrix
- $A \in \mathbb{R}^{d \times d}$ ... orthogonal matrix causing rotation and reflection
- $b \in \mathbb{R}^d$ ... translation factor

⇒ new sum of squared distances is

$$R^2 = \sum_{i=1}^{n} (\tilde{y}_i - s A^T \tilde{x}_i - b)^T (\tilde{y}_i - s A^T \tilde{x}_i - b)$$

→ by seeking the minimal R^2 we can estimate the optimal s, A and b

⇒ Procrustes sum of squares := minimal R^2
↳ measure of "match" between X and Y

⇒ we can plot point-wise residuals between the reference configuration and the final transformed configuration

the arrow shows the distance between $\tilde{y}_i$ and $\tilde{x}_i$

- t-SNE = Stochastic Neighborhood Embedding
  - recent and popular approach for dimension reduction
  - idea: distances can be converted into probabilities
  → similarity between x_i and x_j := probability $p_{j|i}$ that x_i would pick x_j as its neighbour if neighbours were picked in proportion to their probability density under a Gaussian centered at x_i

$$p_{j|i} := \exp\left(-\frac{1}{2}\left(\frac{\|x_i - x_j\|}{\sigma}\right)^2\right) \Big/ \sum_{k \neq i} \exp\left(-\frac{1}{2}\left(\frac{\|x_i - x_k\|}{\sigma}\right)^2\right)$$

Note: $p_{j|i} \neq p_{i|j}$ ⇒ we use $p_{ij} = \frac{1}{2}(p_{j|i} + p_{i|j})$

→ we want to represent high-dim X with low-dim Y

→ we will represent the similarity between y_j and y_i using prob. density q_{ij}

→ assume that the similarities in the low-dim space are governed by a Student-t distribution with one degree of freedom, resulting in

$$q_{ij} = (1 + \|y_i - y_j\|)^2 \Big/ \sum_{k \neq l} (1 + \|y_k - y_l\|)^2$$

↳ to find optimal Y

→ we want to minimize a loss function based on p_{ij} and q_{ij}

⇒ the one which is used is

$$S = \sum_{i \neq j} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right) \quad \dots \quad \frac{p_{ij}}{q_{ij}} =: r \Rightarrow \begin{cases} r = 1 & \Rightarrow \; +0 \\ r < 0 & \Rightarrow \; -\varepsilon \\ r > 0 & \Rightarrow \; +\varepsilon \end{cases}$$

→ S can be minimized using an adaptive learning algorithm

! we should also specify the variance of the Gaussian σ

→ in practice, perplexity is specified instead

↳ it is a function of σ and is interpreted as a smooth measure of the effective number of neighbors considered

↳ it is typically set to be somewhere between 5 and 50

↳ results are sensitive to the choice and should be checked

• Logistic Regression

→ binary classification

→ we have some data and the answer is Yes / No

↳ predicting the presence / absence of a health condition

↳ assessing the likelihood of treatment success

↳ determining the likelihood of a customer to purchase a product

↳ spam detection

⇒ logistic regression gives us the probabilities of Yes and No

⇒ similar to LDA and QDA it is a parametric technique making distributional assumptions over the data

→ motivational example:

→ 92 subjects recorded: Resting Pulse (Low/High), Smokes (Yes/No), Weight (lbs)

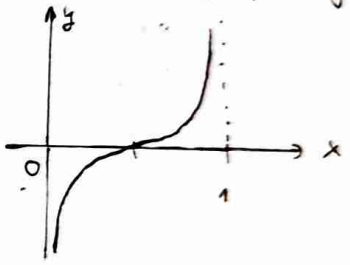
2 lbs ≈ 1 kg

⇒ we want to classify whether the person has Low or High resting pulse

↳ we want $P(\text{Low}) = f(\text{Smokes}, \text{Weight})$ where $\text{Smokes} \in \{0,1\}$, $\text{Weight} \in \mathbb{R}^+$

Def: We define the logit function as

$$\text{logit} : [0,1] \to \mathbb{R}, \quad x \mapsto \ln\left(\frac{x}{1-x}\right)$$



→ Consider the following model

$$\text{logit}(P(\text{Low})) = \alpha + \beta_S \cdot \text{Smokes} + \beta_W \cdot \text{Weight}$$

$$\Rightarrow \frac{P}{1-P} = \exp(\alpha + \beta_S \cdot S + \beta_W \cdot W) \quad \Rightarrow P(\text{Low}) = \frac{\exp(\alpha + \beta_S \cdot S + \beta_W \cdot W)}{1 + \exp(\alpha + \beta_S \cdot S + \beta_W \cdot W)}$$

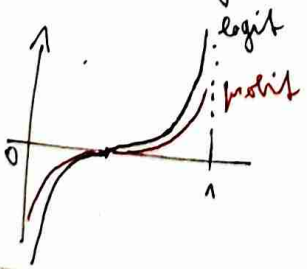
Note: Economists often use the probit model:

$$\Phi^{-1}(z) = w \quad \equiv \quad x \sim N(0,1), \text{ then } P(x < w) = z$$

... Φ = cumulative distr. function of $N(0,1)$

$$\to \Phi^{-1}(P(\text{Low})) = \alpha + \beta_S \cdot S + \beta_W \cdot W$$

! Probit grows slower than logit



⇒ if we fix $\alpha + \beta_S \cdot S + \beta_W \cdot W =: Q$ and consider

$\text{logit}(P(\text{Low})) = \text{probit}(P(\text{Low})) = Q$, then clearly

$P_{\text{logit}} < P_{\text{probit}}$ ⇒ Logit is more strict

↳ requires more evidence for high P

⇒ Consider n independent observations $\tilde{X}_1, \dots \tilde{X}_n$ where

$\tilde{X}_i = (\text{High/Low}, \text{Smokes}_i, \text{Weight}_i)$

↳ $y_i := 1$ if Low, 0 if High

⇒ let's say that the resulting values are $L, L, H, L, H, H, H, \dots L)$

↳ we can look at the probability

$$P(\tilde{X}_1, \dots \tilde{X}_n) = P(L, L, H, L, H, H, H, \dots L) = P(L_1)P(L_2)P(H_3)\cdots P(L_n) \quad \text{, note } P(H_i) = 1 - P(L_i)$$

$$= \prod_{i=1}^{n} \left(\frac{\exp(\alpha + \beta_S S_i + \beta_W W_i)}{1 - \exp(\alpha + \beta_S S_i + \beta_W W_i)} \right)^{y_i} \cdot \prod_{i=1}^{n} \left(\frac{1}{1 - \exp(\alpha + \beta_S S_i + \beta_W W_i)} \right)^{1-y_i}$$

→ we seek α, β_S, β_W that maximize this probability

- best α, β_S, β_W = maximum likelihood estimates
- best $P(\tilde{X}_1, \dots \tilde{X}_n)$ = likelihood

→ if we ask R to do this for us, we get

	Estimate	Std. dev.	z-value	Pr(>z)	
α	−1.99	1.68	−1.18	0.24	
β_S	−1.19	0.55	−2.16	0.03	*
β_W	0.025	0.012	2.04	0.04	*

} statistically significant

→ we also need to consider the std. dev. of the estimates

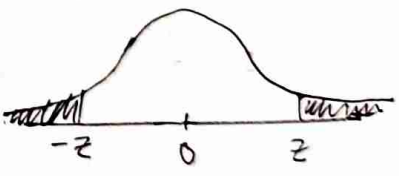
$$\underline{\text{z-value}} = \frac{\text{estimate}}{\text{std. dev}} = \text{relative std. dev} \Rightarrow \text{larger } |z| = \text{more significant}$$

→ to say if the estimate has statistical value, we consider the test

H_0: real value $= 0$ → under H_0: $\frac{\text{estimate} - \mu}{\sigma} \approx \frac{\text{estimate}}{\hat{\sigma}} = \text{z-value} \sim N(0,1)$

H_1: real value $\neq 0$

⇒ we consider the Pr of z being this large



⇒ $\underline{\Pr(>z) = \text{[shaded]} + \text{[shaded]} = p\text{-value of the experiment}}$

↳ for β_S we have $p = 0.03 < 0.05$ ⇒ statistically significant

↳ for α we have $p = 0.24$ ⇒ we aren't very sure about the true α

<u>Interpretation</u>: $\text{logit}(P(\text{low})) = \alpha + \beta_S \cdot \text{Smokes} + \beta_W \cdot \text{Weight}$

- Smokes = True ⇒ P(low) smaller ⇒ higher pulse rate
- ↑ weight ⇒ P(low) larger ⇒ lower pulse rate

→ the standard errors tend to decrease as sample size increases

↳ we can use them to construct 95% confidence intervals for the estimates

$$95\%\ CI = (\text{Estimate}) \pm 2(\text{Std err})$$

$\Rightarrow \beta_S \in (-2.30, -0.09)$ and $\beta_W \in (0.001, 0.05)$

$\Rightarrow \exp(\beta_S) \in (0.1, 0.9)$ and $\exp(\beta_W) \in (1.00, 1.05)$

- **Interactions**

→ if the effect that weight has upon resting pulse would differ depending on whether or not the individual smoked, then we might consider the model

$$\text{logit}(P(\text{Cov})) = \alpha + \beta_S \cdot \text{Smokes} + \beta_W \cdot \text{Weight} + \beta_{SW} \cdot \text{Smokes} \cdot \text{Weight}$$ ← interaction

→ when appropriate, interactions can greatly increase the models performance

- **Akaike's information criterion**

→ choosing a model ~ balancing two opposite goals

- model fit ... how good the $P(\tilde{x}_1, \dots \tilde{x}_n) = \text{likelihood}$ is
- model complexity ... $\rho :=$ # parameters

$\Rightarrow$ $AIC := -2 \log(\text{Likelihood}) + 2\rho$

↳ the model with minimal AIC is the best compromise

- **Logistic regression × LDA**

→ logistic regression can be used to classify data to two clusters C_1 and C_2

$$\text{logit}(P(\tilde{x} \in C_1 | \tilde{x}) = \frac{P(\tilde{x} \in C_1 | \tilde{x})}{P(\tilde{x} \in C_2 | \tilde{x})} = \alpha + \beta^T \tilde{x}$$

→ compare this to LDA

$$\frac{P(\tilde{x} \in C_1 | \tilde{x})}{P(\tilde{x} \in C_2 | \tilde{x})} = \ln \frac{\pi_1}{\pi_2} + \ln \frac{f_1(\tilde{x})}{f_2(\tilde{x})} = 0$$

→ the models have the same form

• Deviance

- another measure for determining a models quality

→ consider this: data points $\tilde{x}_1, \dots \tilde{x}_n \in \mathbb{R}^d$

$$\tilde{x}_j = (\underbrace{x_1, x_2, \dots x_{d-1}}_{\text{covariate vector}}, y) \quad \hookrightarrow \text{classification} \in \{0,1\}$$

→ let $M :=$ # distinct covariate vectors
$n_i :=$ # data points with covariate vector i
$r_i :=$ # data points with covariate vector i & assigned to class 1

→ we want to create a <u>full (saturated) model</u> with M parameters

- if $M = n$, then we can simply let j^{th} parameter I_j to be an indicator for $\tilde{x}_j$
- if $M < n$, let $\pi_i := P[\text{class } 1 \mid \text{covariate vector } i]$
  ↳ based on observed data $\hat{\pi}_i = \frac{r_i}{n_i}$

⇒ then the likelyhood of observed data is

$$L[\text{Data} \mid \pi_1, \dots \pi_M] \propto \prod_{i=1}^{M} \pi_i^{r_i} (1-\pi_i)^{n_i - r_i}$$

$$\ell[\text{Data} \mid \pi_1, \dots \pi_M] = \text{const} + r_i \log(\hat{\pi}_i) + (n_i - r_i)\log(1-\hat{\pi}_i)$$

(log likelihood)

→ ℓ is maximized when $\pi_i = \frac{r_i}{n_i}$, (assign $0 \cdot \log 0 = 0$ if $r_i = 0$)

⇒ let L_{max} be the max. value of L
↳ this represents the <u>best likelihood under any model</u>

→ assume we propose a simpler model
⇒ let L_{mod} be the max. likelihood for this model

<u>Def:</u> The <u>deviance</u> for this model is $\text{Dev} = 2[\log(L_{max}) - \log(L_{mod})]$

⇒ the <u>smaller the deviance</u>, the <u>better the model</u>

→ it is also sometimes called the <u>residual deviance</u>

<u>Def:</u> Now consider the simplest model with just 1 parameter based on the number of data points assigned to class 1 ... $\pi = \frac{\#\text{ in } 1}{n}$

↳ the deviance for this model is the <u>null deviance</u> $= 2[\log L_{max} - \log L_{null}]$

<u>Note:</u> $L_{null} \leq L_{mod} \leq L_{max}$

- lets assume H_0: The proposed model is true

↳ its can be shown that under H_0, the deviance is approximately distributed as a χ^2_{M-k-1} distribution where k = # of parameters of the model

⇒ if the deviance is greater than the 95% point of the appropriate chi-squared distribution, then H_0 probably isn't true and the model is false

⇒ we can not assign a prob. to H_0 holding true, but we can question it

How to do this in R?

calculate: 1 - pchisq(residual_deviance, deg_of_freedom)

↳ this gives us the p-value of the experiment

→ when this value is very small ($\leq 5\%$) then there is statistical evidence that there is a significant diff. between the model of interest and the saturated model

→ we might also want to check against the null deviance to see if the model is better than nothing

→ the smaller the M and larger n, the more trustworthy this test is