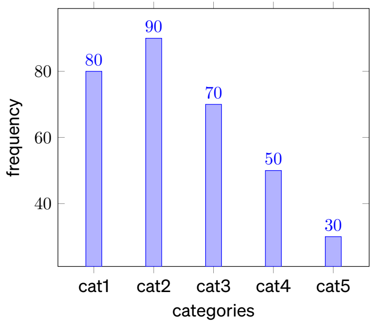
POLIMI GRADUATE SCHOOL OF MANAGEMENT

EXPLORATORY DATA ANALYSIS

PERCORSO EXECUTIVE DATA SCIENCE AND BUSINESS ANALYTICS

Mauricio Soto - mauricioabel.soto@polimi.it

GRAPHICAL ANALYSIS CATEGORICAL ATTRIBUTE

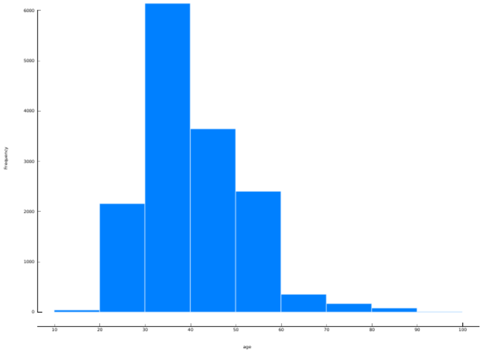


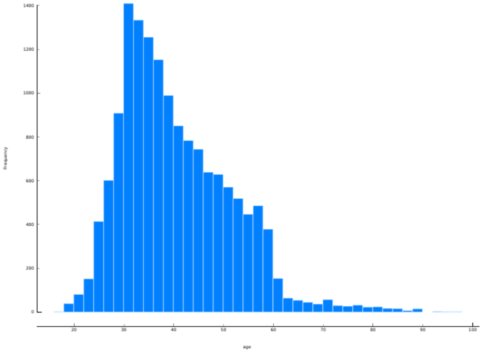
$$f_h = \frac{e_h}{m} = \frac{card\{i \in \mathcal{M} : x_i = cat_h\}}{m}$$

for large samples

$$f_h \approx P(x = cat_h)$$

GRAPHICAL ANALYSIS NUMERICAL ATTRIBUTE





POLIMI GRADUATE SCHOOL OF MANAGEMENT

CENTRAL TENDENCY

- Mean:

$$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

- Median:

$$x^{\text{med}} = x_{(m+1)/2}, \qquad x^{\text{med}} = (x_{m/2} + x_{m/2+1})/2$$

- Mode
- Midrange:

$$x^{\text{midr}} = (x_{\text{max}} + x_{\text{min}})/2$$

- Geometric mean:

$$\bar{\mu}_{\text{geom}} = \sqrt[m]{\prod_{i}^{m} x_i}$$

MEASURE OF DISPERSION - NUMERICAL

- Range:

$$x_{\text{max}} - x_{\text{min}}$$

- Mean absolute deviation :

$$MAD = \frac{1}{m}\sum_{i=1}^{m} |x_i - \bar{\mu}|$$

- Sample variance:

$$\bar{\sigma}^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{\mu})^2$$

- Sample standard deviation :

$$\bar{\sigma} = \sqrt{\bar{\sigma}^2}$$

- Coefficient of Variation:

$$CV = 100\frac{\bar{\sigma}}{\mu}$$

MEASURE OF DISPERSION - CATEGORICAL

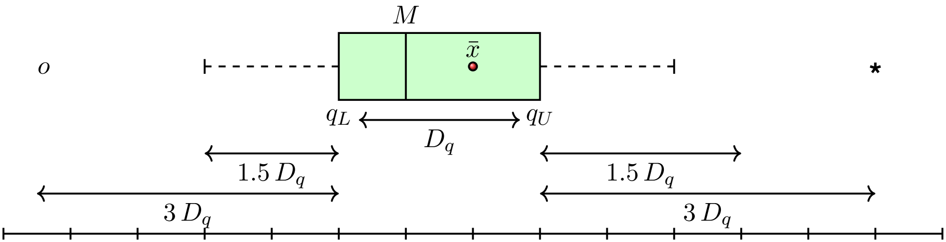
- **Gini index:**

$$Gini = 1 - \sum_{h=1}^{H} f_h^2 \qquad \in [0, (H-1)/H]$$

- **Entropy index:**

$$Entropy = -\sum_{h=1}^{H} f_h \log_2 f_h \qquad \in [0, \log_2 H]$$

BOX-PLOT



- Interquartile range $D_q = q_U - q_L = q_{0.75} - q_{0.25}$
- internal lower edge= $q_L - 1.5\,D_q$
- external lower edge= $q_L - 3\,D_q$

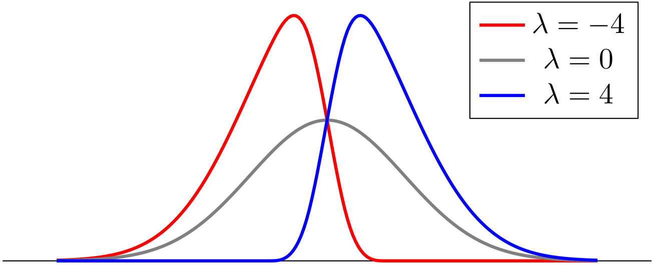
MEASURE RELATIVE LOCATION

- **Mead-Mean**: Mean of values between q_L and q_U
- **Trimmed-Mean**: Mean of values between q_p and $q_{(1-p)}$
- **Winsorized-Mean**: Map values smaller (bigger) than $q_p(q_{(1-p)})$ to $q_p(q_{(1-p)})$ and then compute the mean

ASYMMETRY

$$\bar{\mu}_3 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{\mu})^3, \qquad \text{Skewness} = I_{as} = \frac{\bar{\mu}_3}{\bar{\sigma}^3}$$

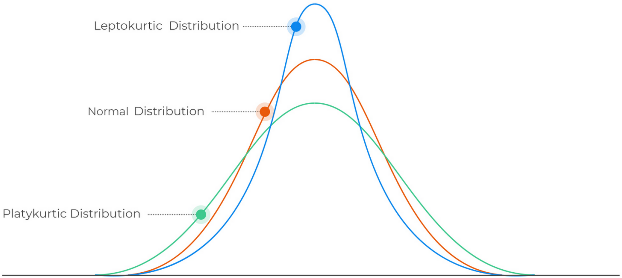
- $I_{as} > 0$ right asymmetry
- $I_{as} < 0$ left asymmetry
- $I_{as} = 0$ symmetric



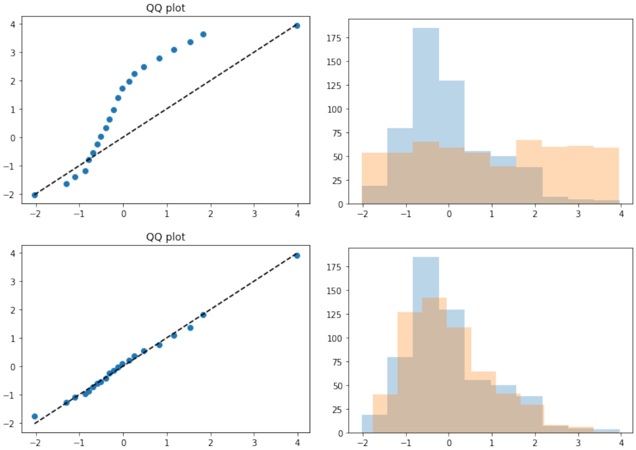
EMPIRICAL DENSITY

$$\bar{\mu}_4 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{\mu})^4, \qquad \text{Kurtosis} = I_{kurt} = \frac{\bar{\mu}_4}{\bar{\sigma}^4} - 3$$

- $I_{kurt} > 0$ Hypernormal
- $I_{kurt} < 0$ Hyponormal
- $I_{kurt} = 0$ Normal



COMPARING DISTRIBUTIONS - QQ-PLOTS



MEASURE OF CORRELATION

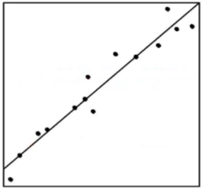
- **Sample covariance:**

$$cov(a_j, a_k) = \frac{1}{m-2} \sum_{i=1}^{m} (x_{ij} - \bar{\mu}_j)(x_{ik} - \bar{\mu}_k)$$

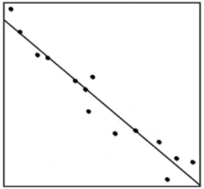
- **Sample LINEAR correlation:**

$$r_{jk} = \frac{cov(a_j, a_k)}{\bar{\sigma}_j \, \bar{\sigma}_j} \qquad \in [-1, 1]$$

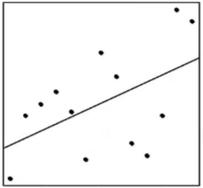
LINEAR CORRELATION



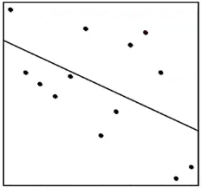
(a) r=0.96



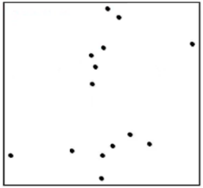
(b) r=-0.96



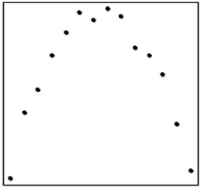
(c) r=0.6



(d) r=-0.6



(e) r=0.0



(f) r=0.0

CORRELATION ON CATEGORICAL ATTRIBUTES

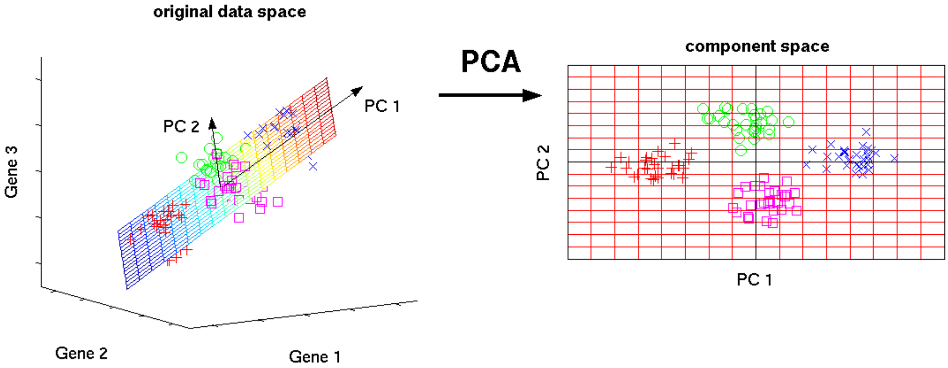
	family		
area	0	1	totale
1	2	4	6 (f_1)
2	4	2	6
3	2	5	7
4	3	3	6
totale	11 (g_1)	14 (g_2)	25

Two attributes are independent if

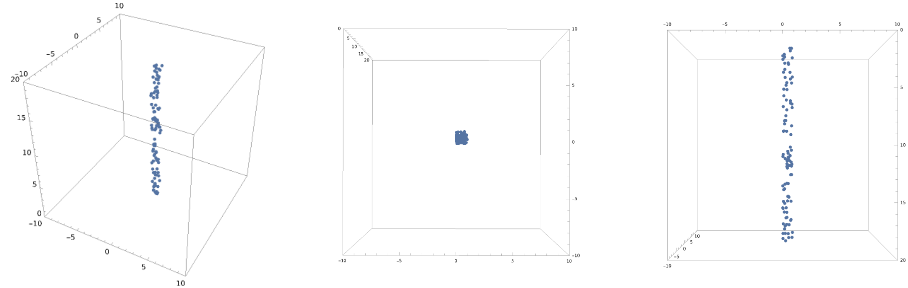
$$\frac{t_{r1}}{g_1} = \frac{t_{r2}}{g_2} \qquad r = 1, 2, \ldots, J$$

Contingency tables

PCA: PRINCIPAL COMPONENT ANALYSIS



PCA-INTUITION



Compute projections that better capture the **variance**

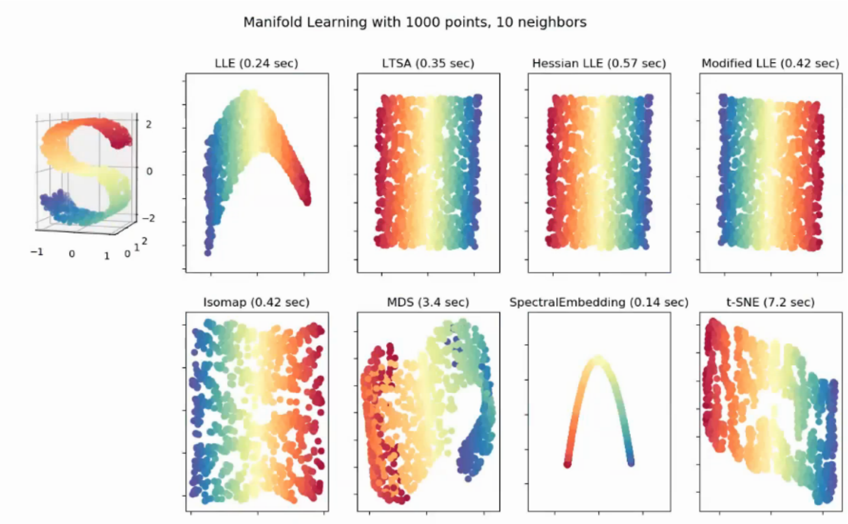
PCA: PRINCIPAL COMPONENT ANALYSIS

- Covariance data matrix $V = X'X$
- $\bar{x}_{ij} = x_{ij} - \bar{\mu}_j$
- New components p_j obtained as a linear transformation of original data
  $p_j = X\, w_j$
- Variance of $p_j = w_j'\, X'X\, w_j = w_j' V w_j$
- Maximizing the variance:

$$\max_{w_1} w_1' V w_1$$
$$\text{s.t. } w_1' w_1 = 1$$

- w_j is the j-th eigenvector of V, which explains a variance λ_j which is the j-th eigenvector

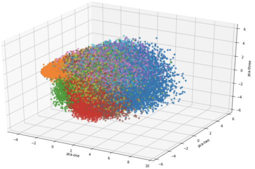
NONLINEAR REDUCTION



T-SNE: T-DISTRIBUTED STOCHASTIC NEIGHBOUR EMBEDDING

We aim to project observations preserving observation distance.

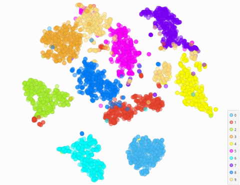
High-Dimensional Space



Distance as Normal Distribution

$$p_{ij} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq l} \exp\left(-\left\|x_k - x_l\right\|^2 / 2\sigma_i^2\right)}$$

Low-Dimensional Space



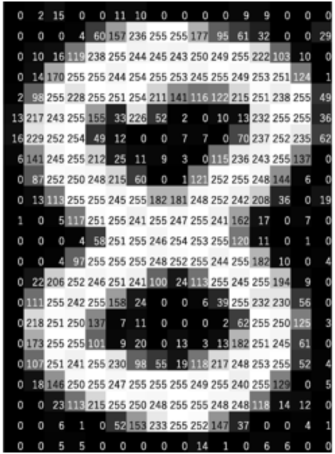
Distance as t-Students Distribution

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

We try to minimize the divergence between the distributions

$$KL(P\|Q) = \sum_i \sum_{j \neq i} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right) \qquad \text{(Kullback-Leibler)}$$

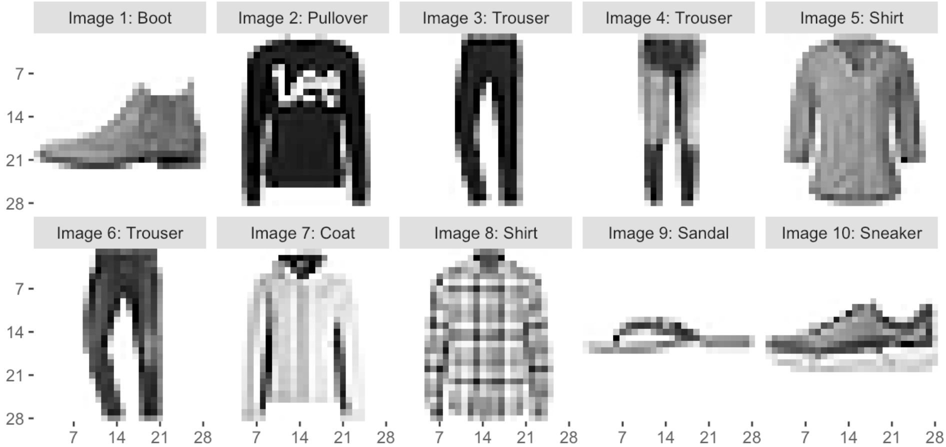
EJEMPLO B - MNIST DATASET



0	2	15	0	0	11	10	0	0	0	0	9	9	0	0	0
0	0	0	4	60	157	236	255	255	177	95	61	32	0	0	29
0	10	16	119	238	255	244	245	243	250	249	255	222	103	10	0
0	14	170	255	255	244	254	255	253	245	255	249	253	251	124	1
2	98	255	228	255	251	254	211	141	116	122	215	251	238	255	49
13	217	243	255	155	33	226	52	2	0	10	13	232	255	255	36
16	229	252	254	49	12	0	0	7	7	0	70	237	252	235	62
6	141	245	255	212	25	11	9	3	0	115	236	243	255	137	0
0	87	252	250	248	215	60	0	1	121	252	255	248	144	6	0
0	13	113	255	255	245	255	182	181	248	252	242	208	36	0	19
1	0	5	117	251	255	241	255	247	255	241	162	17	0	7	0
0	0	0	4	58	251	255	246	254	253	255	120	11	0	1	0
0	0	4	97	255	255	255	248	252	255	244	255	182	10	0	4
0	22	206	252	246	251	241	100	24	113	255	245	255	194	9	0
0	111	255	242	255	158	24	0	0	6	39	255	232	230	56	0
0	218	251	250	137	7	11	0	0	0	2	62	255	250	125	3
0	173	255	255	101	9	20	0	13	3	13	182	251	245	61	0
0	107	251	241	255	230	98	55	19	118	217	248	253	255	52	4
0	18	146	250	255	247	255	255	255	249	255	240	255	129	0	5
0	0	23	113	215	255	250	248	255	255	248	248	118	14	12	0
0	0	6	1	0	52	153	233	255	252	147	37	0	0	4	1
0	0	5	5	0	0	0	0	0	14	1	0	6	6	0	0

https://adamharley.com/nn_vis

EJEMPLO C - MNIST FASHION (ZALANDO)



THANK YOU