Correlation measures

Prof. Asim Tewari
IIT Bombay

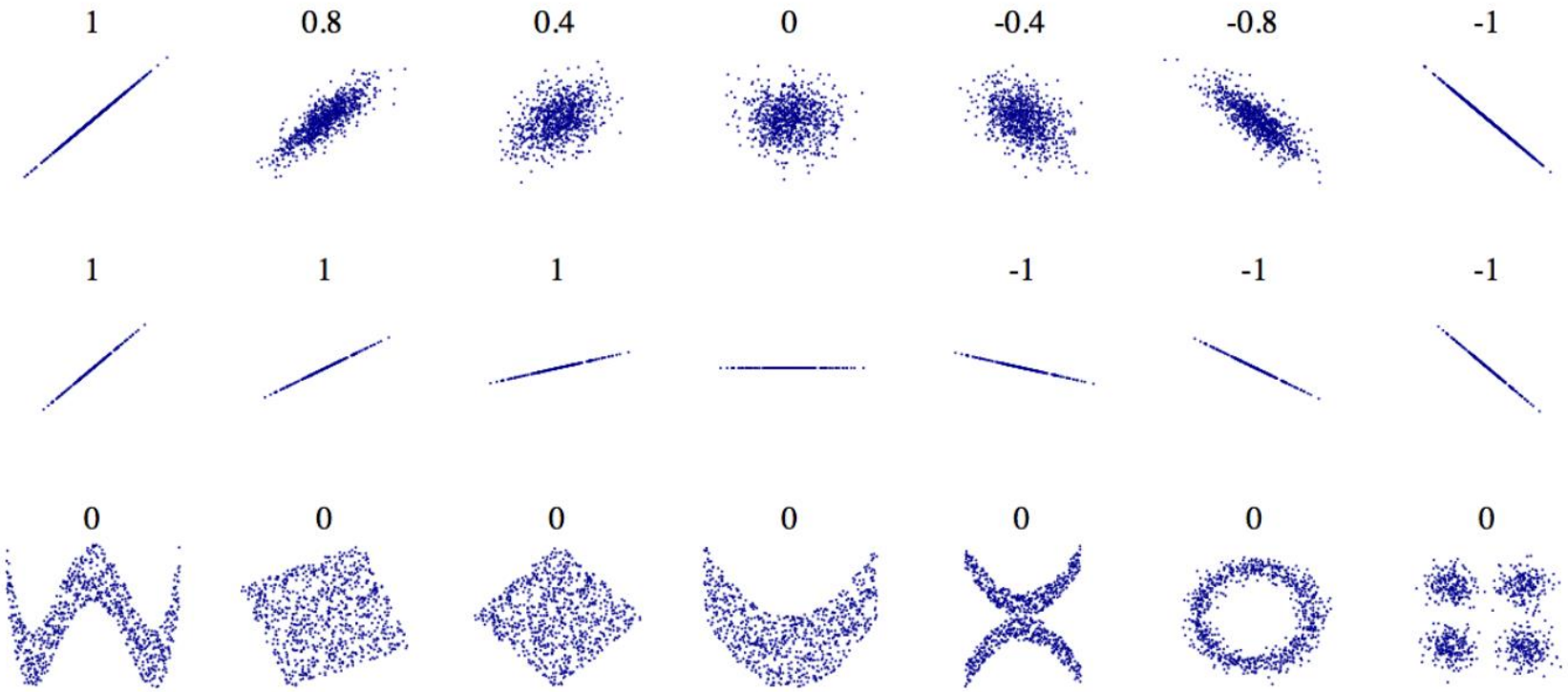
Linear Correlation

Correlation quantifies the relationship between features.

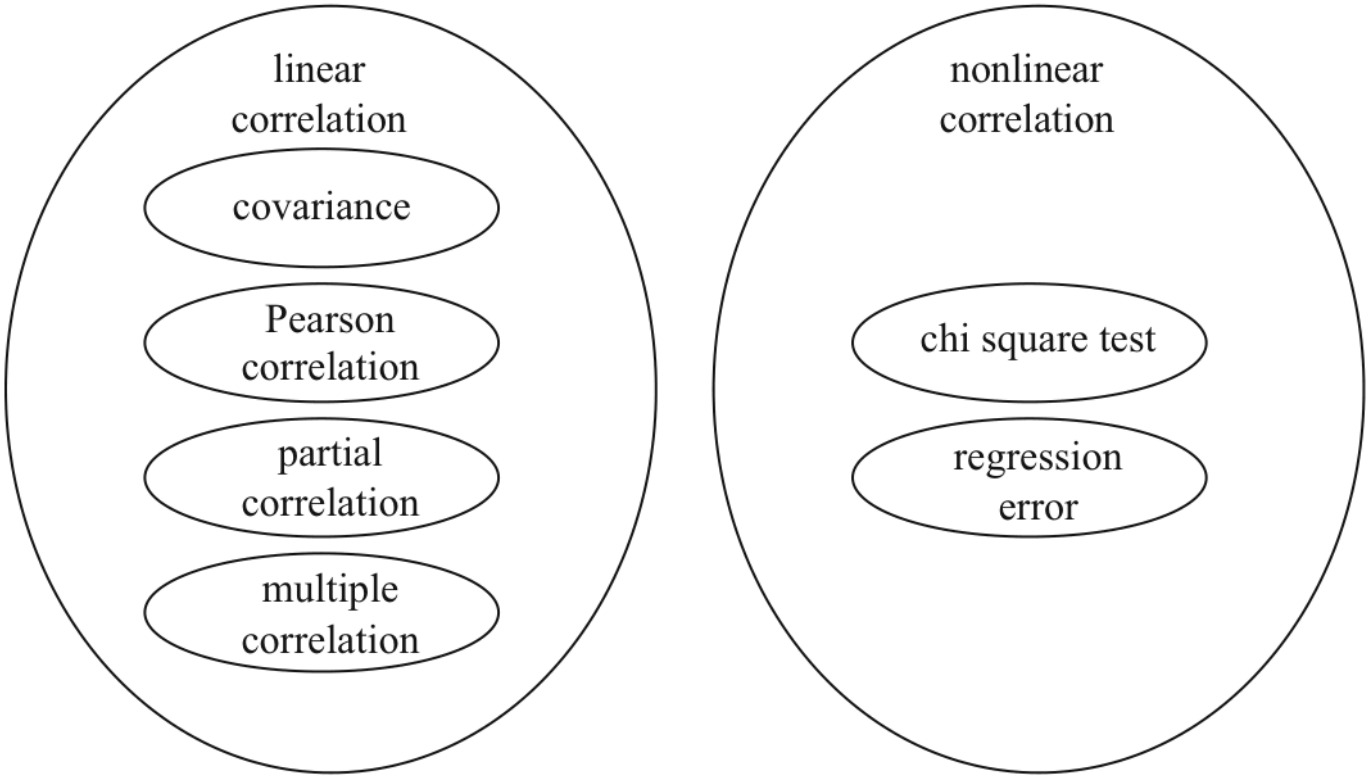
The purpose of correlation analysis is to understand the dependencies between features, so that observed effects can be explained or desired effects can be achieved. For example, for a production plant correlation analysis may yield the features that correlate with the product quality, so the target quality can be achieved by systematically modifying the most relevant features.

$$c_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k^{(i)} - \bar{x}^{(i)})(x_k^{(j)} - \bar{x}^{(j)})$$

Linear Correlation



Some important correlation measures



Pearson correlation coefficient

The Pearson correlation coefficient compensates the effect of constant scaling by dividing the covariance by the product of the standard deviations of both features.

$$s_{ij} = \frac{c_{ij}}{s^{(i)} s^{(j)}}$$

$$= \frac{\sum_{k=1}^{n} (x_k^{(i)} - \bar{x}^{(i)})(x_k^{(j)} - \bar{x}^{(j)})}{\sqrt{\left(\sum_{k=1}^{n} (x_k^{(i)} - \bar{x}^{(i)})^2\right)\left(\sum_{k=1}^{n} (x_k^{(j)} - \bar{x}^{(j)})^2\right)}}$$

$$= \frac{\sum_{k=1}^{n} x_k^{(i)} x_k^{(j)} - n\,\bar{x}^{(i)}\bar{x}^{(j)}}{\sqrt{\left(\sum_{k=1}^{n} \left(x_k^{(i)}\right)^2 - n\left(\bar{x}^{(i)}\right)^2\right)\left(\sum_{k=1}^{n} \left(x_k^{(j)}\right)^2 - n\left(\bar{x}^{(j)}\right)^2\right)}}$$

Pearson correlation coefficient

$$s_{ij} = \frac{c_{ij}}{\sqrt{c_{ii}c_{jj}}}$$

Notice that for μ–σ standardized data, covariance and Pearson correlation are equal, since

$$c_{ii} = c_{jj} = 1 \quad \Rightarrow \quad s_{ij} = \frac{c_{ij}}{\sqrt{c_{ii}c_{jj}}} = c_{ij}$$

Correlation and Causality

Correlation does not imply causality. A correlation between x and y may indicate four different causal scenarios or a combination of these:

1. Coincidence
2. x causes y
3. y causes x
4. z causes both x and y

Correlation and Causality

Coincidence

Even though the data suggest a correlation, this might be a coincidence, so x and y do not possess any causal connection (scenario 1).

x causes y or y causes x

If there is a causal connection between x and y, then correlation does not distinguish whether x causes y or y causes x (scenarios 2 and 3), for example a correlation between the consumption of diet drinks and obesity does not tell us whether diet drinks cause obesity or obesity causes people to drink diet drinks.

z causes both x and y

Finally, there might be no causal connection between x and y: instead both x and y may be caused by z (scenario 4). For example a correlation between forest fires and harvest does not imply that forest fires increase harvest nor that harvest causes forest fires. Instead, it may be that both forest fires and harvest are caused by sunny weather.

This is called spurious correlation or third cause fallacy. Correlation analysis does not distinguish between these scenarios is valid, so often additional expert knowledge is required.

Correlation and Causality

If $x^{(i)}$ and $x^{(j)}$ are correlated and also both correlated with $x^{(k)}$ (like in the spurious correlation scenario), then we might want to know the correlation between $x^{(i)}$ and $x^{(j)}$ without the influence of $x^{(k)}$. This is called the partial or conditional correlation which is defined as

$$s_{ij|k} = \frac{s_{ij} - s_{ik}s_{jk}}{\sqrt{(1 - s_{ik}^2)(1 - s_{jk}^2)}}$$

Correlation and Causality

The correlation between $x^{(i)}$ and $x^{(j)}$ without the influence of the two features $x^{(k)}$ and $x^{(l)}$ is called bipartial correlation defined as

$$s_{i|k,j|l} = \frac{s_{ij} - s_{ik}s_{jk} - s_{il}s_{jl} + s_{ik}s_{kl}s_{jl}}{\sqrt{(1 - s_{ik}^2)(1 - s_{jl}^2)}}$$

Correlation and Causality

The correlation of $x^{(i)}$ with a whole group of features $x^{(j1)}$,… $x^{(jq)}$ is called multiple correlation and is defined as

$$s_{i,(j_1,\ldots,j_q)} = \sqrt{(s_{ij_1} \ldots s_{ij_q}) \cdot \begin{pmatrix} 1 & s_{j_2j_1} & \ldots & s_{j_1j_q} \\ s_{j_1j_2} & 1 & \ldots & s_{j_2j_q} \\ \vdots & \vdots & \ddots & \vdots \\ s_{j_1j_q} & s_{j_2j_q} & \ldots & 1 \end{pmatrix}^{-1} \cdot \begin{pmatrix} s_{ij_1} \\ s_{ij_2} \\ \vdots \\ s_{ij_q} \end{pmatrix}}$$

Notice that S_{ij}=[-1,1] holds for the (simple) correlation, but not necessarily for the partial, bipartial, and multiple correlations.

Correlation and Causality

$$s_{i,(j_1,\ldots,j_q)} = \sqrt{(s_{ij_1} \ldots s_{ij_q}) \cdot \begin{pmatrix} 1 & s_{j_2j_1} & \ldots & s_{j_1j_q} \\ s_{j_1j_2} & 1 & \ldots & s_{j_2j_q} \\ \vdots & \vdots & \ddots & \vdots \\ s_{j_1j_q} & s_{j_2j_q} & \ldots & 1 \end{pmatrix}^{-1} \cdot \begin{pmatrix} s_{ij_1} \\ s_{ij_2} \\ \vdots \\ s_{ij_q} \end{pmatrix}}$$

For q=1, multiple correlation becomes the simple correlation.

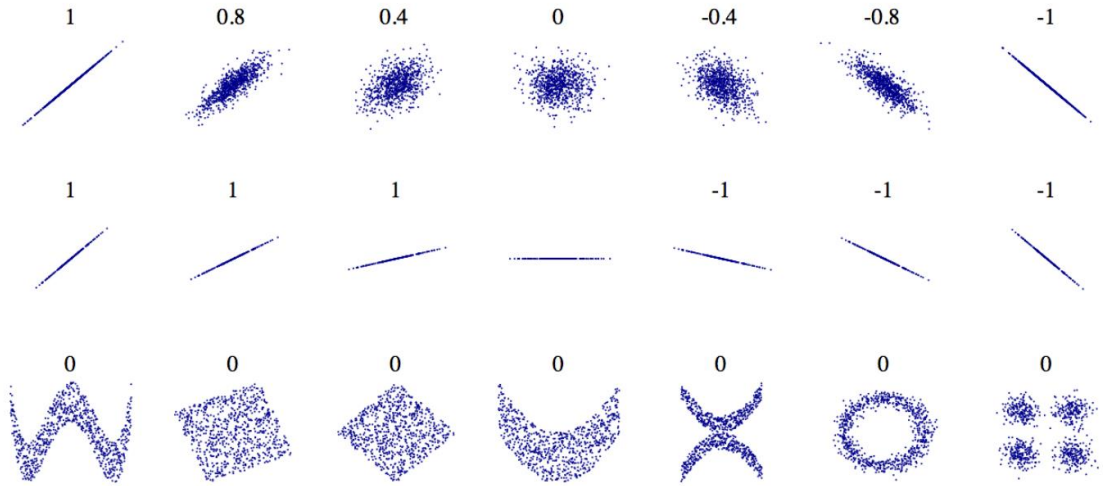
$$s_{i,(j_1)} = |s_{ij_1}|$$

For q=2 we obtain

$$s_{i,(j_1,j_2)} = \sqrt{\frac{s_{ij_1}^2 + s_{ij_2}^2 - 2s_{ij_1}s_{ij_2}s_{j_1j_2}}{1 - s_{j_1j_2}^2}}$$

Chi-Square Test for Independence

- The correlation methods assume linear dependencies between features.
- Strong nonlinear dependencies might yield small or even zero linear correlations.

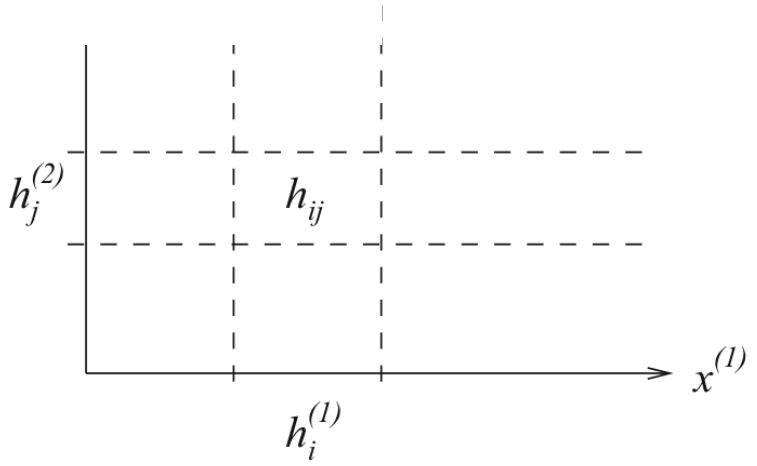


- A method to quantify the nonlinear correlation between features is the Chi-Square test for independence

Chi-Square Test for Independence

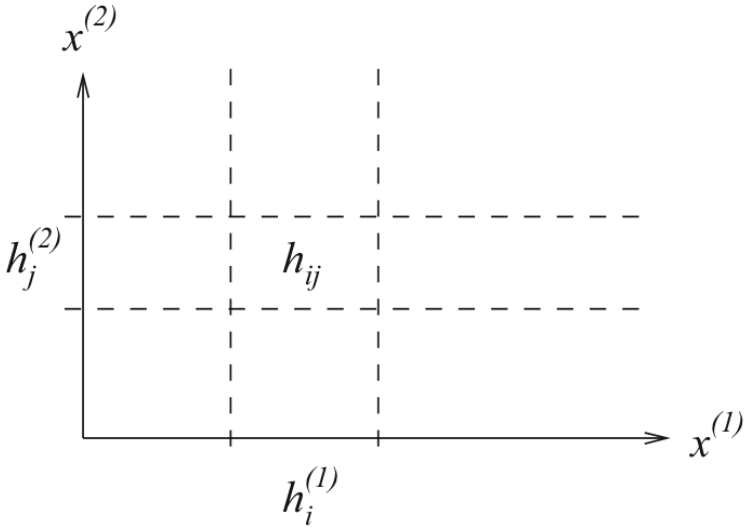
To quantify the nonlinear correlation between two continuous features $x^{(1)}$ and $x^{(2)}$, we first compute the histograms of $x^{(1)}$ and $x^{(2)}$ for r and s bins, respectively.

$$h^{(1)} = (h_1^{(1)}, \ldots, h_r^{(1)}), \quad h^{(2)} = (h_1^{(2)}, \ldots, h_s^{(2)})$$



Chi-Square Test for Independence

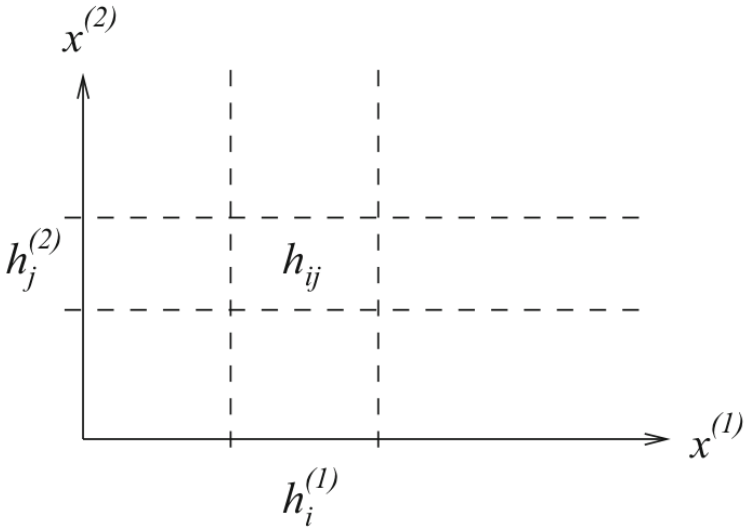
Then we count the number of data that fall into each combination of the ith bin of $x^{(1)}$, I = 1,..,r, and the jth bin of $x^{(2)}$, j=1,..,s, denote this count as h_{ij}, and write these counts as a matrix



$$H = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1s} \\ h_{21} & h_{22} & \cdots & h_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ h_{r1} & h_{r2} & \cdots & h_{rs} \end{pmatrix}$$

The Figure illustrates the counts $h^{(1)}$, $h^{(2)}$ and hij, i=1,..,r, j=1,..,s. The histograms of $x^{(1)}$ and $x^{(2)}$ are the row and column sums of H, respectively.

Chi-Square Test for Independence



$$\sum_{j=1}^{s} h_{ij} = h_i^{(1)}, \quad i = 1, \ldots, r$$

$$\sum_{i=1}^{r} h_{ij} = h_j^{(2)}, \quad j = 1, \ldots, s$$

If the features are independent, then the probability of a data point falling into the bin combination h_{ij} is equal to the product of the probability of falling into bin $h_i^{(1)}$ and the probability of falling into bin $h_i^{(2)}$, so

$$\frac{h_{ij}}{n} = \frac{h_i^{(1)}}{n} \cdot \frac{h_j^{(2)}}{n} \quad \Rightarrow \quad h_{ij} = \frac{h_i^{(1)} \cdot h_j^{(2)}}{n}$$

where

$$n = \sum_{i=1}^{r} \sum_{j=1}^{s} h_{ij} = \sum_{i=1}^{r} h_i^{(1)} = \sum_{j=1}^{s} h_j^{(2)}$$

Chi-Square Test for Independence

Similar to Sammon's mapping, the deviation of h_{ij} from complete independence can be quantified using the absolute square error

$$E_1 = \left(h_{ij} - \frac{h_i^{(1)} \cdot h_j^{(2)}}{n} \right)^2$$

the relative square error

$$E_2 = \left(h_{ij} - \frac{h_i^{(1)} \cdot h_j^{(2)}}{n} \right)^2 \Big/ \left(\frac{h_i^{(1)} \cdot h_j^{(2)}}{n} \right)^2$$

or a compromise between absolute and relative square error

$$E_3 = \left(h_{ij} - \frac{h_i^{(1)} \cdot h_j^{(2)}}{n} \right)^2 \Big/ \left(\frac{h_i^{(1)} \cdot h_j^{(2)}}{n} \right)$$

Chi-Square Test for Independence

Just as in Sammon's mapping we choose the compromise E_3 and obtain the chi-square test statistic

$$E_3 = \left(h_{ij} - \frac{h_i^{(1)} \cdot h_j^{(2)}}{n} \right)^2 / \left(\frac{h_i^{(1)} \cdot h_j^{(2)}}{n} \right)$$

We obtain the chi-square test statistic as:

$$\chi^2 = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(n \cdot h_{ij} - h_i^{(1)} \cdot h_j^{(2)} \right)^2}{h_i^{(1)} \cdot h_j^{(2)}}$$

Chi-Square Test for Independence

$$\chi^2 = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(n \cdot h_{ij} - h_i^{(1)} \cdot h_j^{(2)} \right)^2}{h_i^{(1)} \cdot h_j^{(2)}}$$

The hypothesis that the features are independent is rejected if

$$\chi^2 > \chi^2(1 - \alpha, r - 1, s - 1)$$

where α is the significance level. Small values of χ^2 confirm the hypothesis that the features are independent. The chi-square distribution is monotonic, so the lower χ^2 the higher the stochastic independence between the considered pair of features. Therefore, to produce a list of pairs of features in order of their nonlinear correlation it is sufficient to sort the plain χ^2 values.

Chi-Square Test for Independence

Just as in Sammon's mapping we choose the compromise E_3 and obtain the chi-square test statistic

$$E_3 = \left(h_{ij} - \frac{h_i^{(1)} \cdot h_j^{(2)}}{n} \right)^2 / \left(\frac{h_i^{(1)} \cdot h_j^{(2)}}{n} \right)$$

We obtain the chi-square test statistic as:

$$\chi^2 = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(n \cdot h_{ij} - h_i^{(1)} \cdot h_j^{(2)} \right)^2}{h_i^{(1)} \cdot h_j^{(2)}}$$