Chapter 5: Correlation

1. Linear Correlation
2. Correlation and Causality
3. Chi Square Test for Independence

Prof. Dr. Thomas A. Runkler

Empirical Covariance Matrix

- wanted:
  quantification of correlation between
  data components $x^{(i)}$ and $x^{(j)}$
- covariance matrix of X

$$
\begin{aligned}
c_{ij} &= \frac{1}{n-1} \sum_{k=1}^{n} (x_k^{(i)} - \bar{x}^{(i)})(x_k^{(j)} - \bar{x}^{(j)}) \\
&= \frac{1}{n-1} \left(\sum_{k=1}^{n} x_k^{(i)} x_k^{(j)} - n\, \bar{x}^{(i)} \bar{x}^{(j)} \right) \quad \in \mathrm{IR}
\end{aligned}
$$

with mean

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k$$

Prof. Dr. Thomas A. Runkler

Pearson's Correlation Coefficient

- large variances of $x^{(i)}$ or $x^{(j)}$ imply large covariances c_{ij}, independent of correlation
- correlation matrix of X

$$s_{ij} = \frac{c_{ij}}{s^{(i)} s^{(j)}} \quad \in [-1, 1]$$

with standard deviation

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

so $s^{(i)} = \sqrt{c_{ii}}$, hence

$$s_{ij} = \frac{c_{ij}}{\sqrt{c_{ii} c_{jj}}} \quad \in [-1, 1]$$

Prof. Dr. Thomas A. Runkler

Pearson's Correlation Coefficient

- correlation matrix of X

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

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

Prof. Dr. Thomas A. Runkler

Correlation and Causality

- A correlation between x and y may indicate
  1. coincidence
  2. x causes y
  3. y causes x
  4. z causes both x and y

- example 3

  drinking diet drinks leads to obesity

- example 4 (spurious correlation / third cause fallacy)

$$\begin{array}{r} \text{forest fires} \sim \text{sunshine} \\ \text{corn yield} \sim \text{sunshine} \\ \hline \Rightarrow \text{forest fire} \sim \text{corn yield} \end{array}$$

Prof. Dr. Thomas A. Runkler

(Bi–)Partial Correlation

- correlation between $x^{(i)}$ and $x^{(j)}$ without influence of $x^{(k)}$

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

- correlation of $x^{(i)}$ and $x^{(j)}$ without influence of $x^{(k)}$ and $x^{(l)}$

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

Multiple Correlation

- correlation between $x^{(i)}$ and the features $x^{(j_1)}, \ldots, x^{(j_q)}$

$$s_{i,(j_1,\ldots,j_q)} = \sqrt{(s_{ij_1} \ldots s_{ij_q}) \cdot \begin{pmatrix} 1 & s_{j_2 j_1} & \ldots & s_{j_1 j_q} \\ s_{j_1 j_2} & 1 & \ldots & s_{j_2 j_q} \\ \vdots & \vdots & \ddots & \vdots \\ s_{j_1 j_q} & s_{j_2 j_q} & \ldots & 1 \end{pmatrix}^{-1} \cdot \begin{pmatrix} s_{ij_1} \\ s_{ij_2} \\ \vdots \\ s_{ij_q} \end{pmatrix}}$$

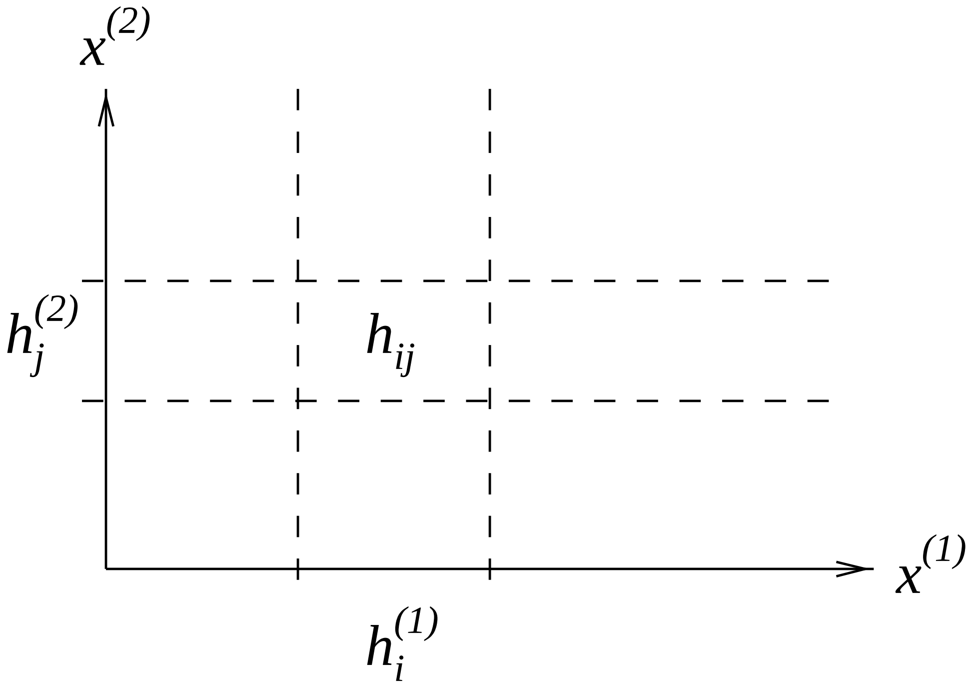
- example $q = 1$

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

- example $q = 2$

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

Chi Square Test for Independence



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

Chi Square Test for Independence

- stochastical independence

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

- mixed error measure

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

- hypothesis of independence is rejected if

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

- monotonicity $\Rightarrow$ χ^2 is measure for nonlinear correlation

