Numerical data description

Part of Exploratory Data Analysis

Location: "where"

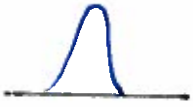
- Sample mean
- Median
- Mode

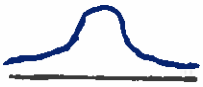
Scale: "spread"

- Sample variance / standard deviation
- Median absolute deviation (MAD)
- Interquartile range (IQR)

Shape: "form"

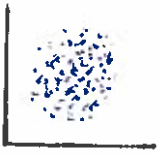
- Sample skewness
- Galton's measure of skewness

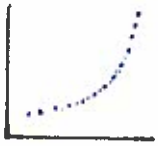




- Sample kurtosis
- Robust kurtosis

Dependence: "relation"





- Sample covariance / correlation
- Kendall's τ

Multivariate measures

Now: d-dim. random vector $\underline{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix}$

n observations

$\rightsquigarrow$ Observation matrix $X = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ x_{21} & \ddots & \\ \vdots & & \\ x_{d1} & & x_{dn} \end{pmatrix}$

x_{ij} = j-th observation of i-th attribute

Attention: here we have $d \times n$ observation matrices !!

In MLPR it was the other way

Both are common in literature

Notation:

$x_{i:}$: i-th row vector of X

$x_{i:} = (x_{i1}, \ldots, x_{in})$

$x_{:j}$: j-th column vector of X

$x_{:j} = \underline{x}_j = \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{d1} \end{pmatrix}$

- Sample covariance matrix:

$$cov(X) = \begin{pmatrix} cov(x_{1:}, x_{1:}) & cov(x_{1:}, x_{2:}) & \cdots & cov(x_{1:}, x_{d:}) \\ \vdots & & \ddots & \vdots \\ cov(x_{d:}, x_{1:}) & \cdots & & cov(x_{d:}, x_{d:}) \end{pmatrix}$$

Estimate of covariance matrix

$$Cov[\underline{x}] = \mathbb{E}\left[(\underline{x} - \mathbb{E}[\underline{x}])(\underline{x} - \mathbb{E}[\underline{x}])^T\right]$$

Symmetric with variances on the diagonal

Positive semi-definite:

$$\underline{w}^T Cov[\underline{x}]\, \underline{w} = \mathbb{E}[\underbrace{\underline{w}^T(\underline{x} - \mathbb{E}[\underline{x}])}_{\text{scalor}}\underbrace{(\underline{x} - \mathbb{E}[\underline{x}])^T \underline{w}}_{\text{scalor}}]$$

$$= \mathbb{E}[(\underline{w}^T(\underline{x} - \mathbb{E}[\underline{x}]))^2] \geq 0$$

Eigenvalue decomposition:

$$Cov[\underline{x}] = U \Lambda U^T$$

where

Λ diagonal matrix with eigenvalues $\lambda_1 \ldots \lambda_d$

U orthonormal matrix with eigenvectors

Total variance of d attributes

= sum of eigenvalues

$$\sum_{i=1}^{d} \mathrm{Var}[X_{i:}] = \mathrm{trace}(\mathrm{Cov}[\underline{X}])$$

$$= \mathrm{trace}(U \Lambda U^T)$$

$$= \mathrm{trace}(\Lambda U^T U)$$

$$= \mathrm{trace}(\Lambda) = \sum_{i=1}^{d} \lambda_i$$

Covariance of linear transformation:

$$\mathrm{Cov}[A\underline{X} + \underline{b}] = A\,\mathrm{Cov}[\underline{X}]\,A^T$$

Data preprocessing

Prepare data for further analysis

- Standardisation: make variables comparable

  Often: normalise to have mean 0 and variance 1

  linear transformation

  → express using matrix operations

Given: $X = (X_{:1} \dots X_{:n}) = (\underline{X}_1 \dots \underline{X}_n)$

Sample mean: d-dim. vector $\bar{\underline{X}} = (\bar{X}_{1:} \dots \bar{X}_{d:})$

with $\bar{X}_{i:} = \frac{1}{n} \sum_{j=1}^{n} X_{ij}$

Centred observations: $\tilde{\underline{X}}_i = \underline{X}_i - \bar{\underline{X}}$

→ matrix $\tilde{X} = (\tilde{\underline{X}}_1 \dots \tilde{\underline{X}}_n)$

$$= (\underline{X}_1 - \bar{\underline{X}} \dots \underline{X}_n - \bar{\underline{X}})$$

$$= X - (\underbrace{\bar{\underline{X}} \dots \bar{\underline{X}}}_{n\text{-times}})$$

Centring matrix : $C_n = I_n - \frac{1}{n} 1_n 1_n^\top$

where I_n = identity matrix $(n \times n)$

and $1_n = (1\,1 \ldots 1)^\top$ $(n \times 1)$

It holds: $\tilde{X} = X C_n$

Proof:

$$\bar{\underline{x}} = \frac{1}{n} \sum_{i=1}^{n} \underline{x}_i = \frac{1}{n} X 1_n$$

$$\bar{\underline{x}} 1_n^\top = \begin{pmatrix} \bar{x}_{1:} \\ \vdots \\ \bar{x}_{d:} \end{pmatrix} (1\,1 \ldots 1) = \begin{pmatrix} \bar{x}_{1:} & \cdots & \bar{x}_{1:} \\ \bar{x}_{2:} & & \vdots \\ \vdots & & \\ \bar{x}_{d:} & \cdots & \bar{x}_{d:} \end{pmatrix} = \underbrace{(\bar{\underline{x}} \ldots \bar{\underline{x}})}_{n\text{-times}}$$

$$\Rightarrow (\bar{\underline{x}} \ldots \bar{\underline{x}}) = X \frac{1}{n} 1_n 1_n^\top$$

$$\tilde{X} = X - (\bar{\underline{x}} \ldots \bar{\underline{x}}) = X - X \frac{1}{n} 1_n 1_n^\top$$

$$= X \left(I_n - \frac{1}{n} 1_n 1_n^\top \right) = X C_n \quad \square$$

Properties:

- Idempotence: $C_n C_n = C_n$ "removing mean twice"
- Multiplication from right removes row mean
- Multiplication from left removes column mean

$$C_n \underline{v} = \begin{pmatrix} v_1 - \bar{v} \\ v_2 - \bar{v} \\ \vdots \\ v_n - \bar{v} \end{pmatrix}$$

Expression of sample covariance matrix

$$\text{cov}(X) = \frac{1}{n}\sum_{i=1}^{n}(\underline{x}_i - \bar{\underline{x}})(\underline{x}_i - \bar{\underline{x}})^T = \frac{1}{n}\sum_{i=1}^{n}\tilde{\underline{x}}_i\tilde{\underline{x}}_i^T$$

$$= \frac{1}{n}\tilde{X}\tilde{X}^T = \frac{1}{n}X C_n X^T$$

Scaling to unit variance:

Recall: $\text{Cov}[A\underline{x} + \underline{b}] = A\,\text{Cov}[\underline{x}]\,A^T$

Set $A = \text{diag}\left(\frac{1}{\text{std}(X)}\right)$

where $\text{diag}(\underline{v})$ = diagonal matrix with $\underline{v}$ on diagonal

and $\text{std}(X) = (\text{std}(X_{1:}), \ldots, \text{std}(X_{d:}))^T$

$\rightsquigarrow$ $\text{cov}(AX)$ has ones on diagonal (i.e. unit variance)

Note: correlations are still present

Outlier detection and removal

Outlier: observation that is unusual compared to others

Could occur due to:

- measurement error ⇝ remove
- wrong assumptions ⇝ keep

Indication: e.g. mean >> median

Often used: Tukey's fences

$[Q_1 - k\,IQR,\ Q_3 + k\,IQR]$

where $k \geq 0$ often $k = 1.5$

Observations outside often labelled as outliers