High-dimensional data analysis

Lecture 2

Multidimensional data

Innopolis University

Fall 2018

Bivariate analysis

- Independent (regular) variable: <u>feature</u> of the data, measured and/or manipulated.
- Dependent (target) variable: <u>label</u> of the data, is expected to depend on independent variables.

- Bivariate analysis: only two variables
  - Both independent
  - One dependent, one independent

Multivariate and high-dimensional analysis

- Multivariate analysis (MVA):
  - More than two variables.
  - Usually many independent with one or no dependent.

- High-dimensional analysis (HDA):
  - Multivariate analysis, when #of features $\approx$ #of samples.
  - Or even #of features $\gg$ #of samples (e.g., genetic data and text).

- The basic methods are <u>not</u> specific to HDA, and come from MVA.

Multivariate data analysis

- Goals:
    - understand the structure in the data;
    - summarize data in simpler ways;
    - find the relationship between parts of the data;
    - make decisions based on the data.

Basic statistical measures

- For single feature:
  - Mean
  - Variance
  - Median
  - Quartiles (1, 10, 25, 50, 75, 90, 99)
- For pairs of features:
  - Correlation
    - Pearson (“classical” correlation)
    - Spearman rank correlation
    - Kendall rank correlation

Data visualization

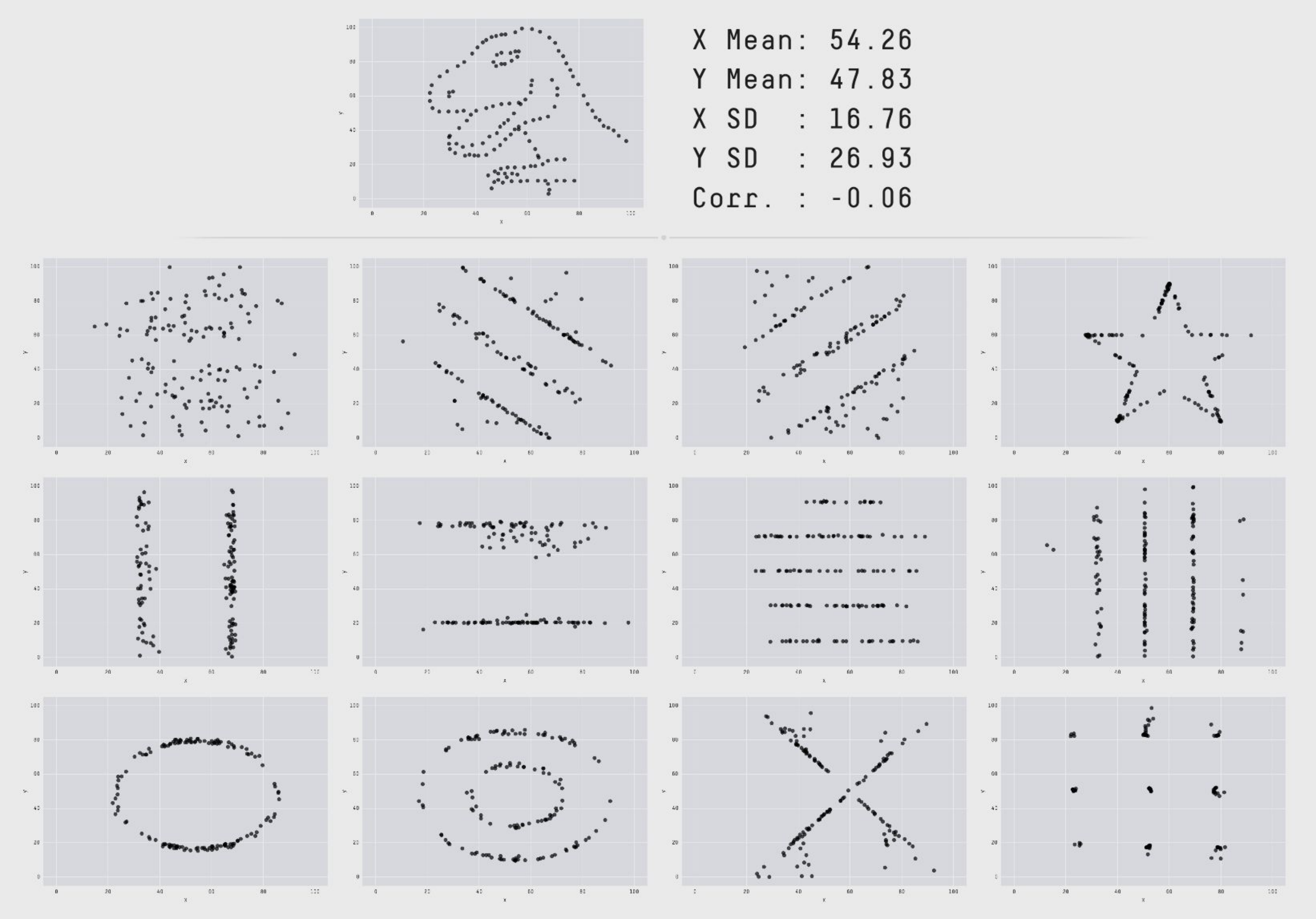
Thinking is more interesting than knowing, but not more interesting than looking at. –Goethe

- There are many statistical measures of data.
- They have limitations:
  - Variance is not meaningful for bi-modal data;
  - mean is skewed by outliers;
  - etc.
- It is way easier to look at your data before running any statistical tests

Data visualization

All these sets have same mean, deviation, and correlation coefficient.

Look at your data!



https://www.autodeskresearch.com/publications/samestats

Data visualization

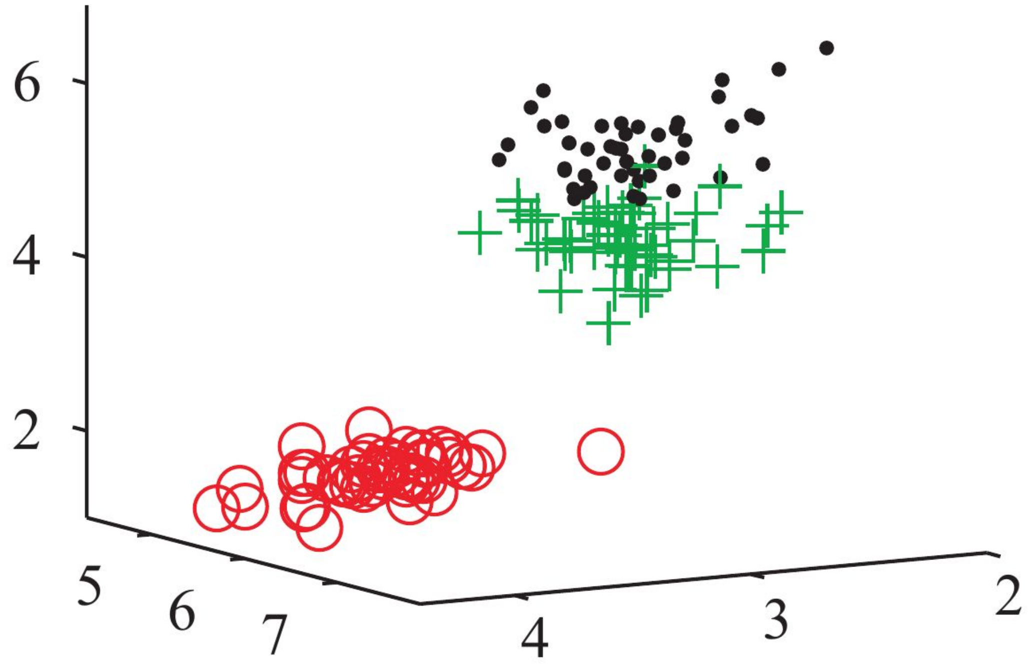
- When we have only two variables, visualization is easy.
- If we have more, we can make pairwise plots.
- But what if we want more variables in single image?

- We discuss only basic approaches to multidimensional data visualization.
- You can look at e.g. https://github.com/d3/d3/wiki/gallery for many-many types of plots.

Three-dimensional plot

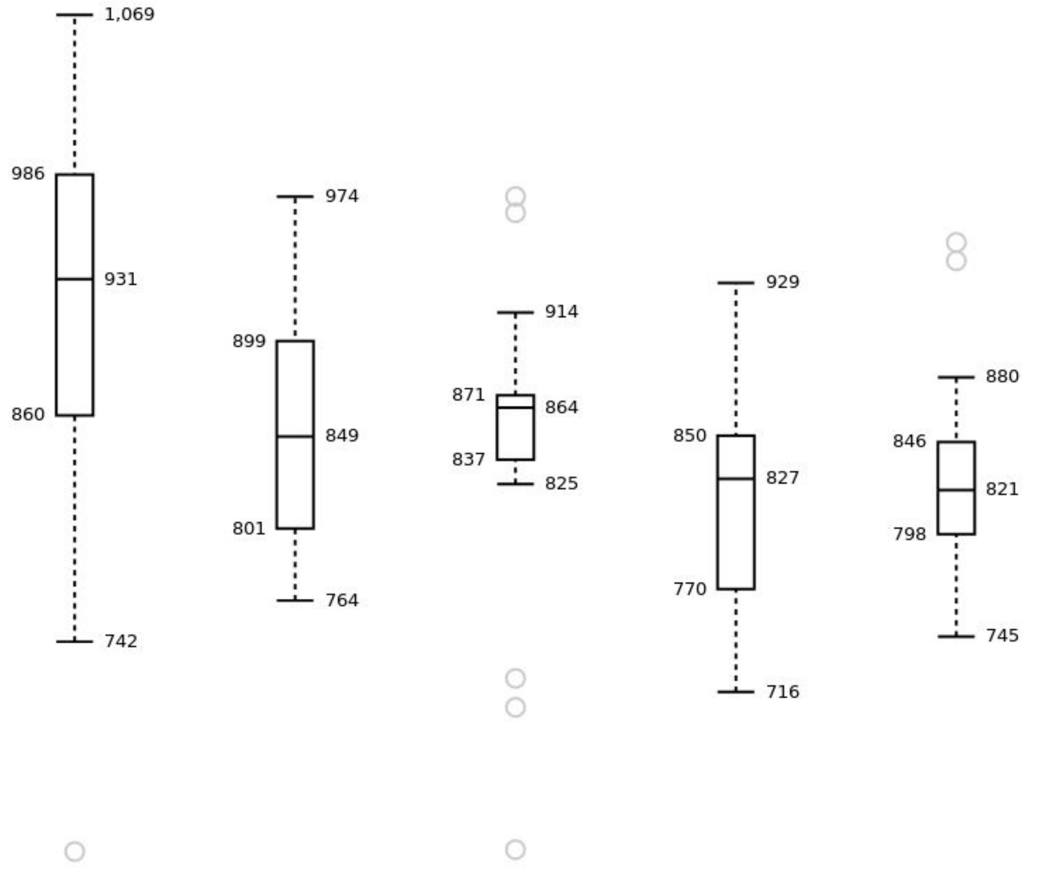
- Projection of 3-dimensional plot on 2D media:
  - Multiple figures
  - Stereo
  - Movie
  - Interactive

- We can also use projections of *n*-dimensional data on 2D plane, but that's impractical



Box-plots

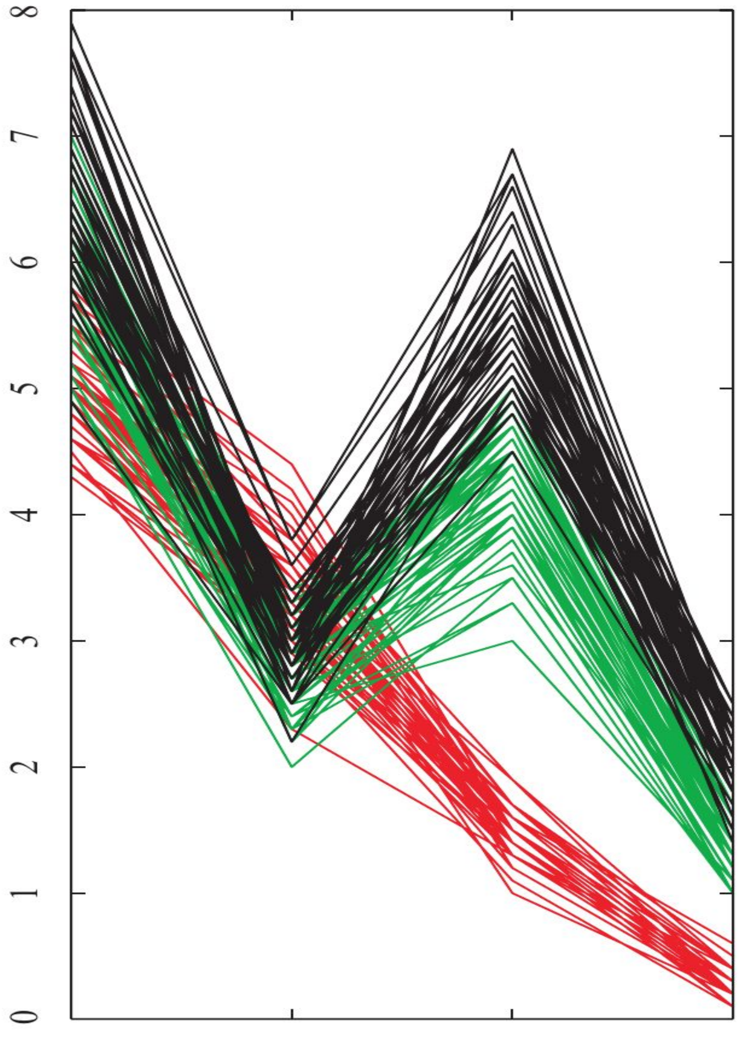
- X-axis: variable
  - we have 5 here
- Y-axis: value
- The box show first, second (median), and third quartiles
- The whiskers usually show Tukey interval
- Values outside Tukey interval are shown explicitly (outliers)



https://bl.ocks.org/mbostock/4061502

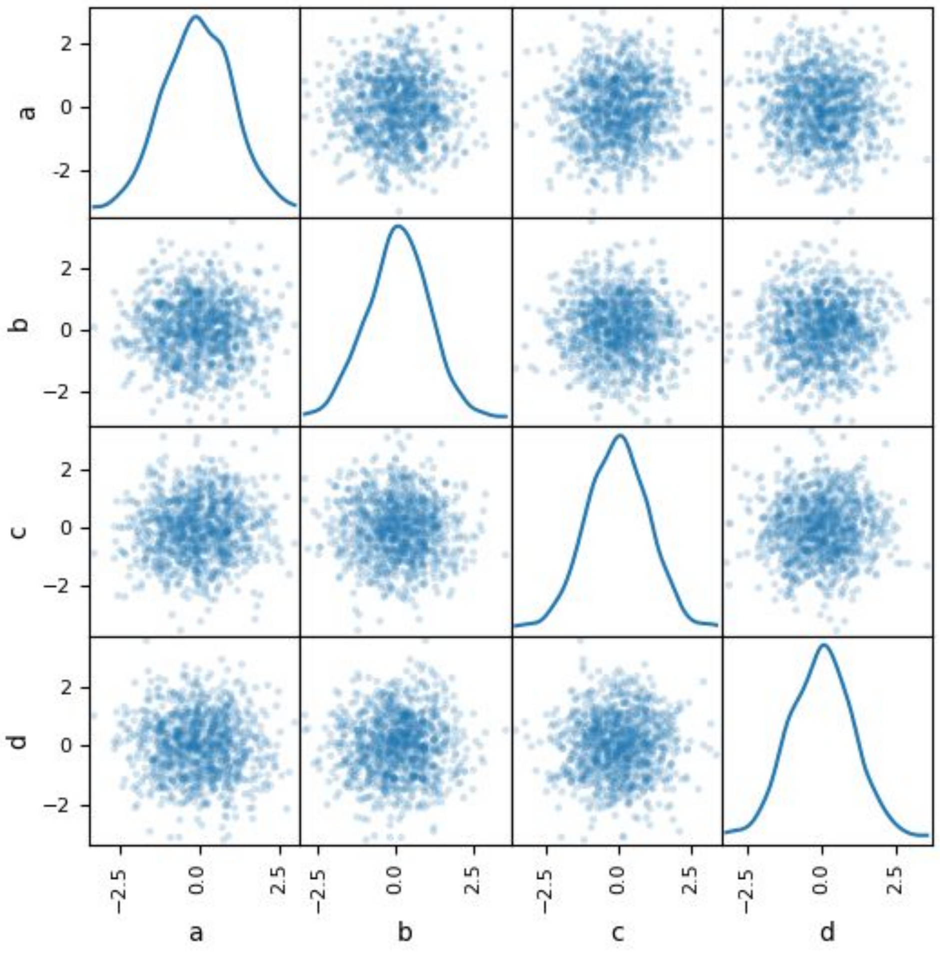
Parallel-coordinate plots

- Like a boxplot, not a timeline
- X-axis: variable
  - we have 4 here
- Y-axis: value
  - all variables are in range [0; 9]
- Each line: single sample
  - Color is label
- Unlike boxplot, can show relationship between variables



Scatter matrix

- Scatter-plots for each pair of variables
- Histogram or kernel density estimation (KDE) on diagonal
- You probably don't want to use it with more than 20 variables



https://pandas.pydata.org/pandas-docs/stable/visualization.html#visualization-scatter-matrix

Random distribution properties

- Probability density function (pdf):

$$f_X(x) = \lim_{\Delta \to 0} P(x \leq X \leq x + \Delta)$$

- Cumulative distribution function (cdf):

$$F_X(x) = P(X < x) = \int_{-\infty}^{x} f_X(t)\, dt$$

Random distribution properties

- Expected value (mean):

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} t\, f_X(t)\, dt = \int_{-\infty}^{+\infty} t\, dF_X(t)$$

- Variance:

$$\mathrm{Var}(X) = \mathbb{E}\left(\left(X - \mathbb{E}(X)\right)^2\right) = \int_{-\infty}^{+\infty} t^2\, dF_X(t) - \mathbb{E}^2(X)$$

- Covariance:

$$\mathrm{Cov}(X,Y) = \mathbb{E}\left(\left(X - \mathbb{E}(X)\right)\left(Y - \mathbb{E}(Y)\right)\right); \quad \mathrm{Cov}(X,X) = \mathrm{Var}(X)$$

Random distribution properties

- Characteristic function:

$$\varphi_X(x) = \mathbb{E}\left(e^{-itX}\right)$$

Multivariate random vectors

- Random vectors – vector-valued functions defined on a sample space.

- In practice we deal with <u>observed</u> data, that we assume has underlying model.
- Observed data is non-random.
- The model might include randomness (e.g., noise).

Population case

- We have model
- We want to find properties of single vector
  - E.g., mean and covariance matrix: $\boldsymbol{X} \sim (\boldsymbol{\mu}, \Sigma)$

random variable

random vector

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_d \end{bmatrix}$$

dimensionality

$$\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{X}) = \begin{bmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_d) \end{bmatrix}$$

$$\Sigma = \mathrm{var}(\boldsymbol{X}) = \begin{bmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) & \cdots & \mathrm{cov}(X_1, X_d) \\ \mathrm{cov}(X_2, X_1) & \mathrm{var}(X_2) & \cdots & \mathrm{cov}(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_d, X_1) & \mathrm{cov}(X_d, X_2) & \cdots & \mathrm{var}(X_d) \end{bmatrix}$$

Population case

- If vector $\boldsymbol{X}$ is d-dimensional (d-variate)
- $\boldsymbol{X} \sim (\boldsymbol{\mu}, \Sigma)$,
- matrix A is $d \times k$, matrix B is $d \times l$,
- then
    1. $A^T\boldsymbol{X} \sim (A^T\boldsymbol{\mu}, A^T\Sigma\mathrm{A})$;
    2. vectors $A^T\boldsymbol{X}$ and $B^T\boldsymbol{X}$ are uncorrelated iff $A^T\Sigma B = \mathbf{0}$

Random sample case

- We have multiple random vectors (samples, measurements)
  - Usually their observed values: $\boldsymbol{X}_i = \boldsymbol{x}_i$
  - $\mathbb{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2\ \cdots\ \boldsymbol{X}_n] = \begin{bmatrix} X_{11} & \cdots & X_{n1} \\ \vdots & \ddots & \vdots \\ X_{1d} & \cdots & X_{nd} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_{\blacksquare 1} \\ \vdots \\ \boldsymbol{X}_{\blacksquare d} \end{bmatrix}$
- We want to find underlying model
  - E.g., find sample mean and covariance: $\mathbb{X} \sim \mathrm{Sam}(\overline{\boldsymbol{X}}, S)$

$$\overline{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i\,; \qquad S = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^T$$

(Univariate) Gaussian distribution

- Two parameters: μ (mean) and σ^2 (variance).

$$X \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

$$f_{\mathcal{N}}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Multivariate Gaussian Distribution

- Random vector $\boldsymbol{X} \in R^k$ is multivariate normally distributed if:
  - $\forall \boldsymbol{v} \in R^k : \boldsymbol{v}^T\boldsymbol{X}$ is normally distributed, and
  - $\exists \boldsymbol{z} \in R^l, \boldsymbol{\mu} \in R^k, A \in R^{k\times l}$, such that
    - every component of $\boldsymbol{z}$ is normally distributed
    - $\boldsymbol{X} = A\boldsymbol{z} + \boldsymbol{\mu}$
  - $\exists \boldsymbol{\mu} \in R^k, \Sigma \in R^{k\times k}$, Σ is symmetric and positive semidefinite, such that
    - the characteristic function of $\boldsymbol{X}$ becomes:
      - $\varphi_{\boldsymbol{X}}(\boldsymbol{x}) = \exp\left(i\boldsymbol{x}^T\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{u}^T\Sigma\boldsymbol{u}\right)$

Random distribution properties

- Probability density function (pdf):

$$f_X(x) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k|\Sigma|}}$$