

# What is multivariate analysis?

- ① First course of statistics: numbers—random variables
- ② **Second course of statistics: vectors of numbers—random vectors**
  - Basis for analysis of more complex objects, e.g. functions, matrices, tensors, images, networks.
- ③ Data Exploration: visualization of relationships between observations.
- ④ Discovering and modeling patterns from dataset: Visualization, Clustering, Multivariate distributions.
- ⑤ Confirming patterns: Inference.
- ⑥ Dimension Reduction: PCA, CCA, SVD.
- ⑦ Predictions: Regression, Classification.

## What is a multivariate dataset?

Multivariate statistical analysis concerns multivariate data where each observation consisting of many measurements on the same subject. We suppose the dataset  $\mathbf{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$  has  $n$  observations (Here,  $n$  is called the sample size), and each observation  $\mathbf{X}_i = (x_{i1}, \dots, x_{ip})$  is a vector in  $\mathbb{R}^p$  (Here,  $p$  is called the dimension). These are often recorded in a  $p \times n$  matrix:

$$\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n) = \begin{pmatrix} x_{11} & \cdots & x_{n1} \\ \vdots & \ddots & \vdots \\ x_{1p} & \cdots & x_{np} \end{pmatrix}$$

# Data Exploration - Visualization

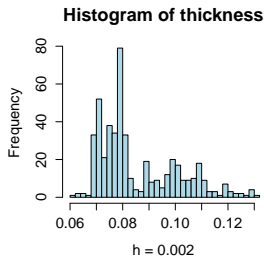
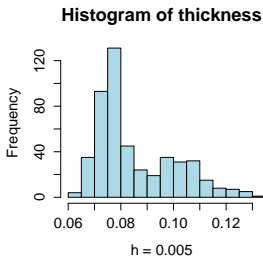
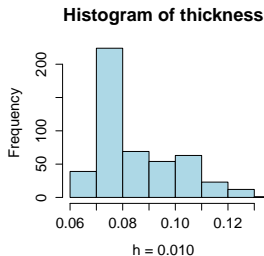
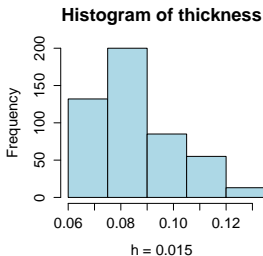
## 1D example – Hidalgo Stamp Data

- $n = 485$  observed values of thickness for Mexico stamps
- over  $> 70$  years
- During 1980s
- Stamp papers produced in several factories?
- No records. Can we guess by looking at the data?

Izenman and Sommer (1988), here we use data “stamp” in R package BSDA.

# Data Exploration - 1D Hidalgo Stamp

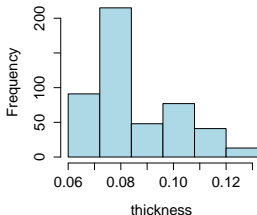
Histograms with different bin widths. How many factories?



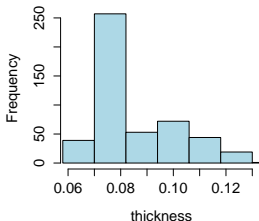
# Data Exploration - 1D Hidalgo Stamp

Histograms with different bin locations. (Fixed bin width 0.012)

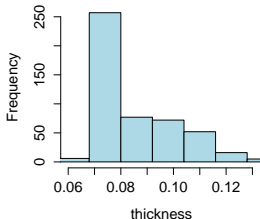
**Histogram of thickness**



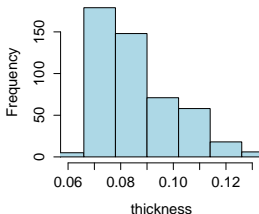
**Histogram of thickness**



**Histogram of thickness**



**Histogram of thickness**



## Data Exploration - Kernel density estimate

- Histograms are dependent on binwidth and bin location.
- Smaller binwidth preferred for exploration.
- Different bin locations can obscure important underlying patterns; A solution is to average out the effect on different bin location “Averaged histogram”
- A more elegant solution is kernel density estimate.
- For  $X_1, \dots, X_n \sim \text{i.i.d. } f(x)$  (continuous pdf), a kernel density estimator of  $f$  is obtained as

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x - X_i}{h}\right),$$

where the kernel  $K(\cdot)$  is a function satisfying  $\int K(x)dx = 1$ .

## Kernel density estimate - illustration

- Toy data:  $x_i = 3, 5, 9, 11, 12, 14$ .
- Consider a kernel density estimate with Gaussian kernel

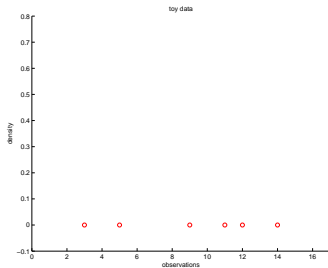
$$K(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2},$$

with bandwidth  $h = 1$ .

- For each  $i$ ,

$$\frac{1}{h} K\left(\frac{x - x_i}{h}\right) = \frac{1}{\sqrt{2\pi}h^2} e^{-\frac{(x-x_i)^2}{2h^2}}.$$

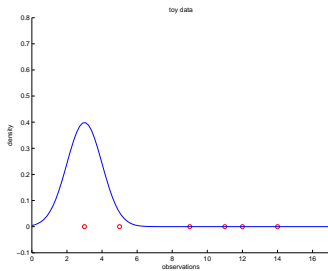
## Kernel density estimate - illustration



- Toy data:  $x_i = 3, 5, 9, 11, 12, 14$ .



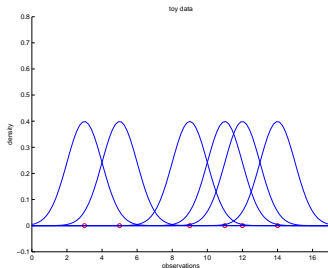
## Kernel density estimate - illustration



- overlaid is, for  $h = 1$ ,

$$\frac{1}{h} K\left(\frac{x - x_1}{h}\right) = \frac{1}{\sqrt{2\pi h^2}} e^{-\frac{(x - x_1)^2}{2h^2}}.$$

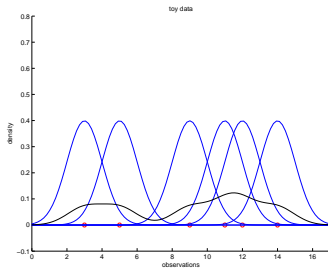
## Kernel density estimate - illustration



- For all  $i = 1, \dots, 6$ ,

$$\frac{1}{h} K\left(\frac{x - x_i}{h}\right) = \frac{1}{\sqrt{2\pi}h} e^{-\frac{(x-x_i)^2}{2h^2}}.$$

## Kernel density estimate - illustration

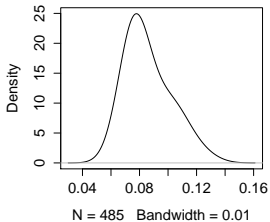


- Kernel Density Estimate (KDE) of the pdf:

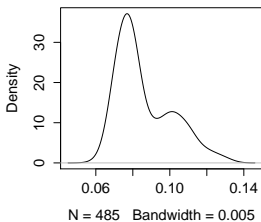
$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x - X_i}{h}\right),$$

# KDE for Hidalgo Stamp

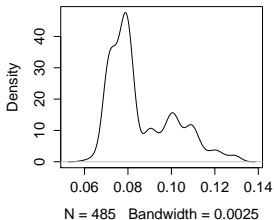
**KDE**



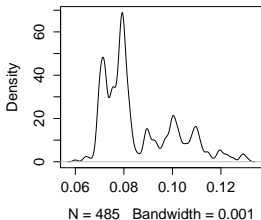
**KDE**



**KDE**



**KDE**

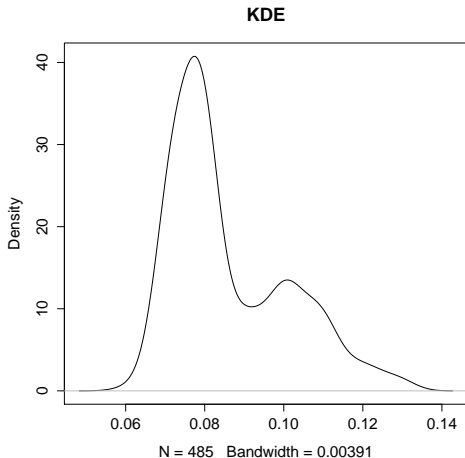


## KDE for Hidalgo Stamp

- Small bandwidth leads undersmooth; Large bandwidth leads oversmooth
- Is it unimodal? Bi-modal? Or several modes? Which modes are really there?
- Choice of bandwidth
  - Important in practice.
  - Controversial issue.
  - Many recommendations (Silverman's rule of thumb, Sheather-Jones Plug-In.)
  - The consensus is that there is never a consensus.

## KDE for Hidalgo Stamp

Bandwidth, by Silverman's rule-of-thumb, is  $(\text{sample size})^{-1/5} \times 90\%$  of minimum of two population standard deviation estimates:  
i) the sample standard deviation, ii)  $IQR/1.34$



# Data Exploration - Multivariate data

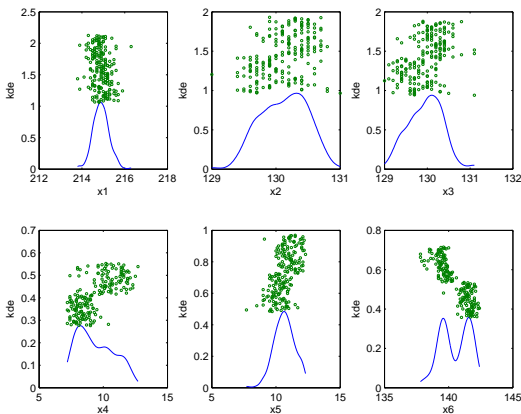
## Dimension $p = 6$ example – Swiss bank notes

- $n = 200$  Swiss bank notes (See Fig. 1.1, Härdle and Simar)
- Each note (obs.) has  $p = 6$  measurements (variables).
- Additional information: first half are genuine; the other half are counterfeit.
- Visualization of 6-dim'l data?
- Can use 6 KDEs overlaid with jitterplot for each measurements (variables)

## Swiss bank notes - Marginal KDEs

Marginal KDEs overlaid with jitterplot for each of 6 variables.

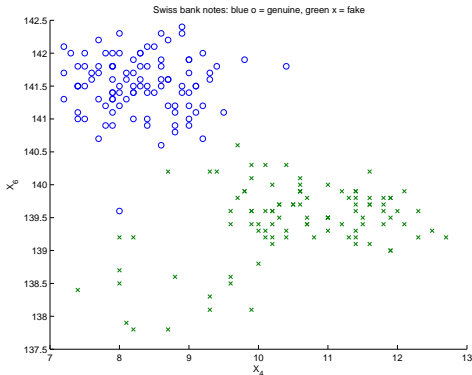
- Informative, realistic when  $p$  is small
- No information about association between variables.





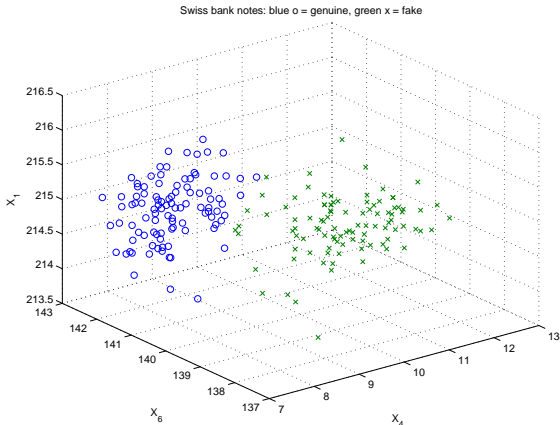
## Swiss bank notes - Scatterplot

- Pairs of variables are best visualized by scatterplot, e.g.  $X_4$  vs  $X_6$  below.
- Understood as point clouds (representing the empirical distributions)
- $\binom{6}{2}$  many pairs to choose from.



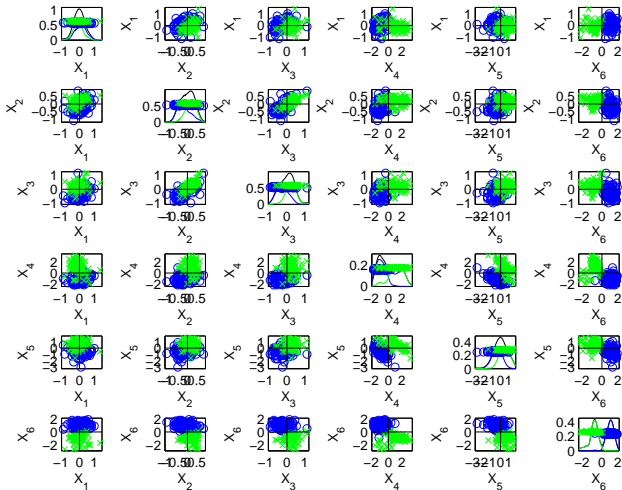
## Swiss bank notes - Scatterplot

- Scatters of three variables can also be informative, if software allows to rotate the axes.
- Otherwise, the 3D scatterplot is a scatterplot of linear combinations of the three variables.



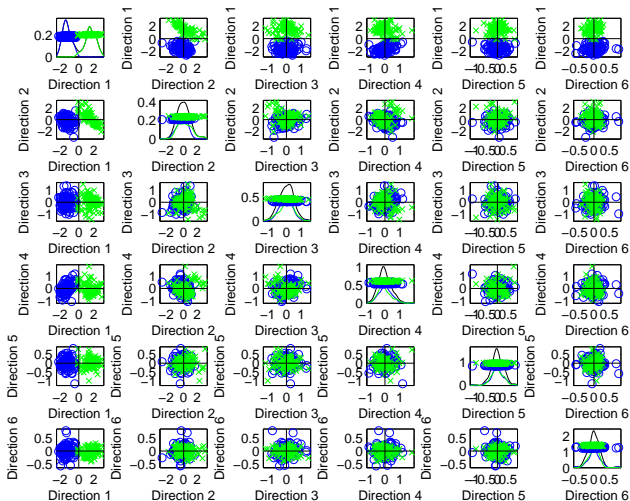
## Swiss bank notes - scatterplot matrix

- A traditional, yet powerful, tool is to construct a matrix of scatterplots. - Too busy with  $p = 6$ .



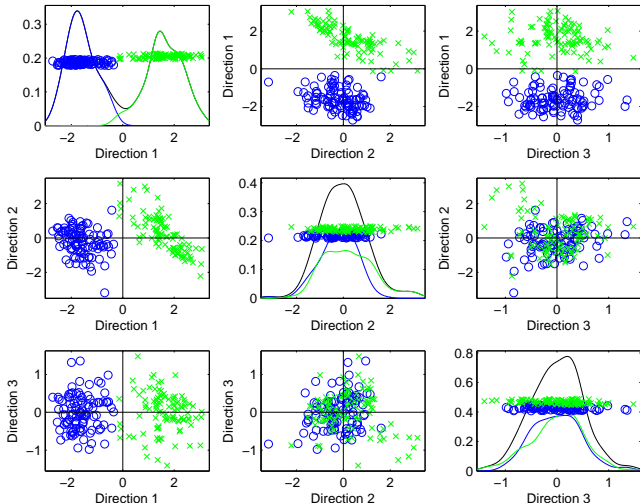
# Swiss bank notes - scatterplot matrix

- Better to visualize with principal component scores.



## Swiss bank notes - scatterplot matrix

- With principal component scores, we can focus on fewer combinations; Dimension Reduction



## Swiss bank notes - related methods

- 1 If obtaining succinct representation of data is of interest, then using the first two principal component scores appeared in the first  $2 \times 2$  block of the scatterplot matrix would do the job (Principal Component Analysis)
- 2 Parametric modeling: Are distributions of Swiss bank note measurements normal (Gaussian)? (Multivariate Normal Distribution)
- 3 Without the information on genuine and counterfeit notes, can we classify  $n = 200$  notes into two distinct groups? (Clustering)
- 4 With the information on genuine and counterfeit notes,
  - Are the means and covariances of genuine and counterfeit notes different? (Statistical inference)
  - When there is a new bank note, how to predict whether the new note is genuine? (Classification)

# Modern challenges

## High dimensional data

- 1 Gene expression data in Golub et al. (1999)
- 2  $p = 7129$  gene expression levels (numeric) for  $n_1 = 47$  subjects with acute lymphoblastic leukemia (ALL) and  $n_2 = 25$  subjects with acute myeloid leukemia (AML).
- 3 Scientific task is to identify new cancer class and/or to assign tumors to known classes (ALL or AML).

### REPORTS

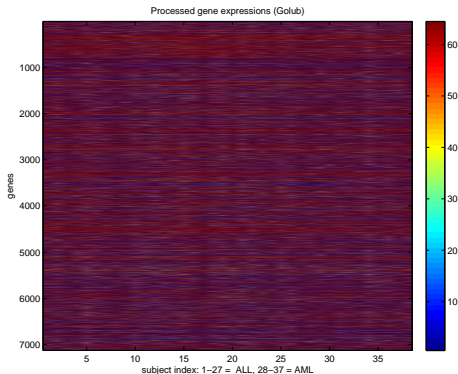
#### **Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring**

T. R. Golub,<sup>1,2\*</sup> D. K. Slonim,<sup>1†</sup> P. Tamayo,<sup>1</sup> C. Huard,<sup>1</sup>  
M. Gaasenbeek,<sup>1</sup> J. P. Mesirov,<sup>1</sup> H. Coller,<sup>1</sup> M. L. Loh,<sup>2</sup>  
J. R. Downing,<sup>3</sup> M. A. Caligiuri,<sup>4</sup> C. D. Bloomfield,<sup>4</sup>  
E. S. Lander<sup>1,5\*</sup>

## Golub data

Taking a subgroup of data with 27 ALLs and 11 AMLs.

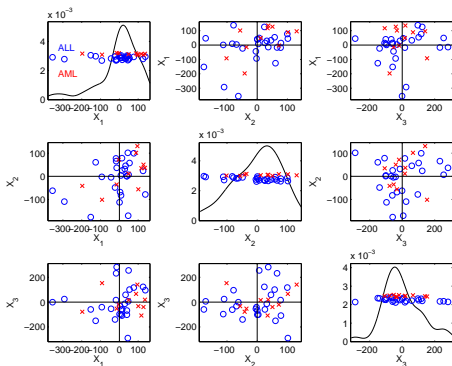
A visualization of the matrix  $X_{p \times n}$ . Is it helpful?





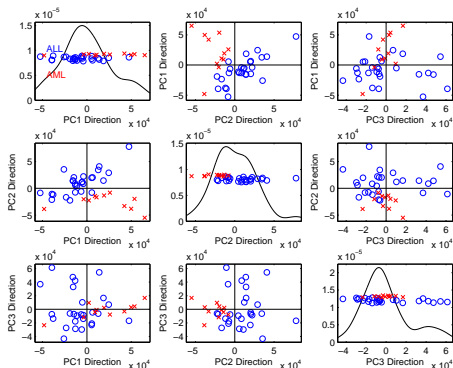
## Golub data

Scatterplot matrix for the first three genes (variables).



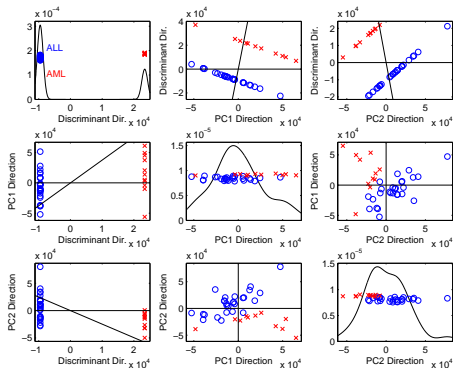
## Golub data

Scatterplot matrix using the first three Principal Component Scores. A pattern there?



## Golub data

Scatterplot matrix using the a lineiar combination of all variables (that leads a good separation of two groups) and two Principal Component Scores. Better pattern?



Next: review on matrix algebra.