Lecture 2: Visualization

MAST30034 Applied Data Science

Dr. Karim Seghouane
School of Mathematics & Statistics
The University of Melbourne

August 1, 2023

Outline

- Data Science and Applied Data Science Aims
- Examples of Pattern Visualization
- Random Vectors and Properties
- Histograms
- Spectral Decomposition
- Notions of Simulation
- Optimism and Performance Evaluation

What is Data Science about?

Data science deals mostly with the development and use of unsupervised learning methods which aim to look for and find

- interesting patterns and structure,
- relationships or features in data

which we can and will be further exploited in making decision about a current situation or predicting quantities for new data. Essential are

- Machine Learning: computational issues and algorithmic developments and
- Statistical Learning: statistical methods and associated foundations required in the analysis of the data.

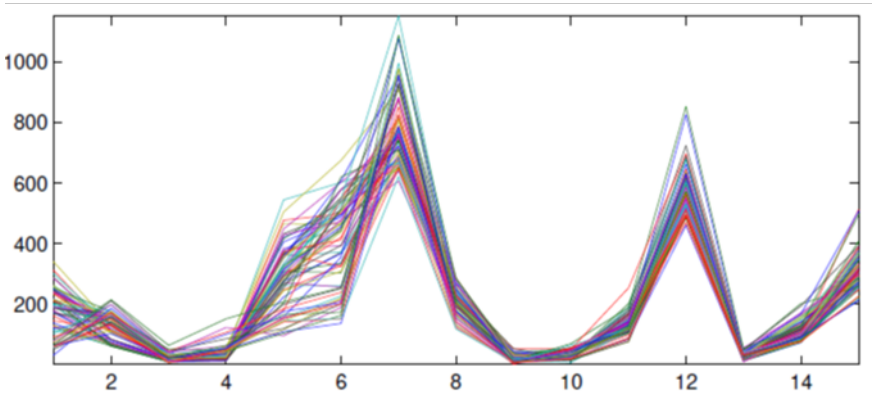
What is Applied Data Science about?

In applied data science we focus on learning about and applying statistical methods for

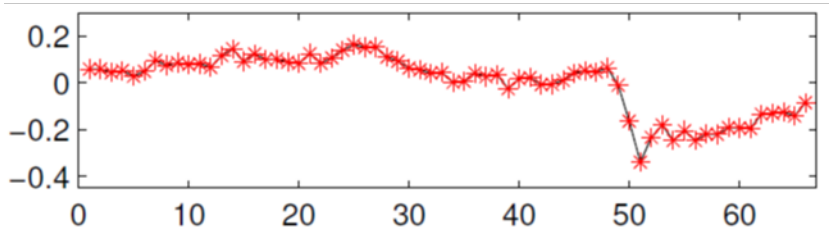
- uncovering patterns and structure, and unveiling relevant features and information in data;
- finding features which allow a meaningful division of the observations into groups with similar properties; and
- constructing rules for making decisions and predicting new outcomes based on available data.

Pattern Visualization in Data

The illicit drug market market data: time series over 66 months
Parallel coordinate plot



can detect the heroin shortage in early 2001: month 49



Pattern Visualization in Data

HIV data - how many clusters are in HIV^+ and HIV^- data ?
Scatterplots

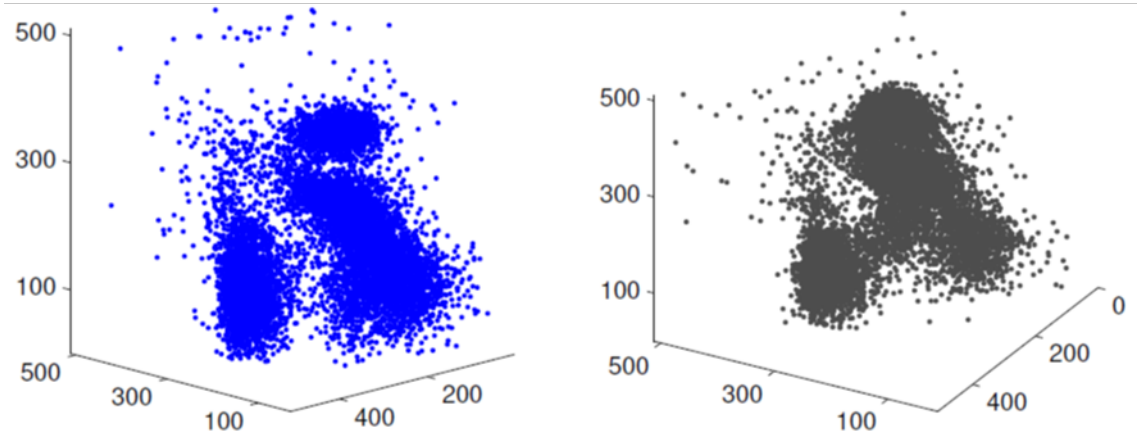
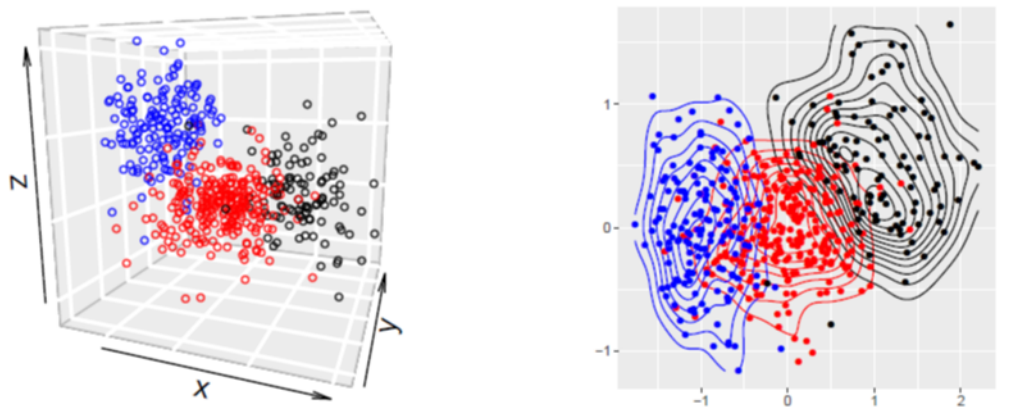


Figure: HIV^+ data (left) and HIV^- data (right) with variables CD3, CD8 and CD4

- What are the distinguishing features between the data sets ?
- How are the clusters arranged ?

Pattern Visualization in Data

Learn from data and classify current and new data; and find a rule that divides data - which variables does it depend on ?



Pattern Visualization in Data

- Identifying the level of activation.

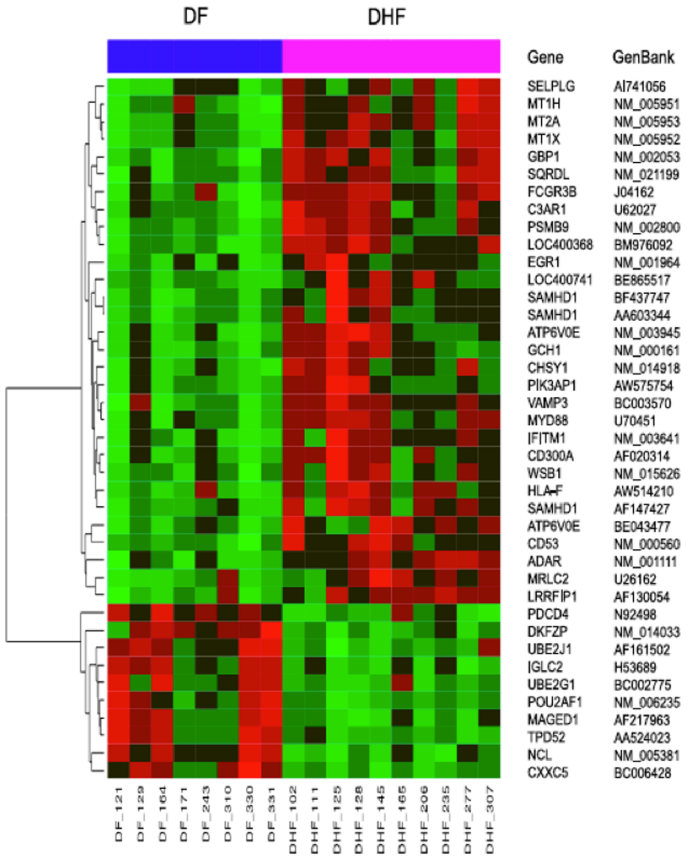


Figure: Heatmap for the gene expression microarray data matrix. Red and green code for high and low expression values, respectively.

Pattern Visualization in Data

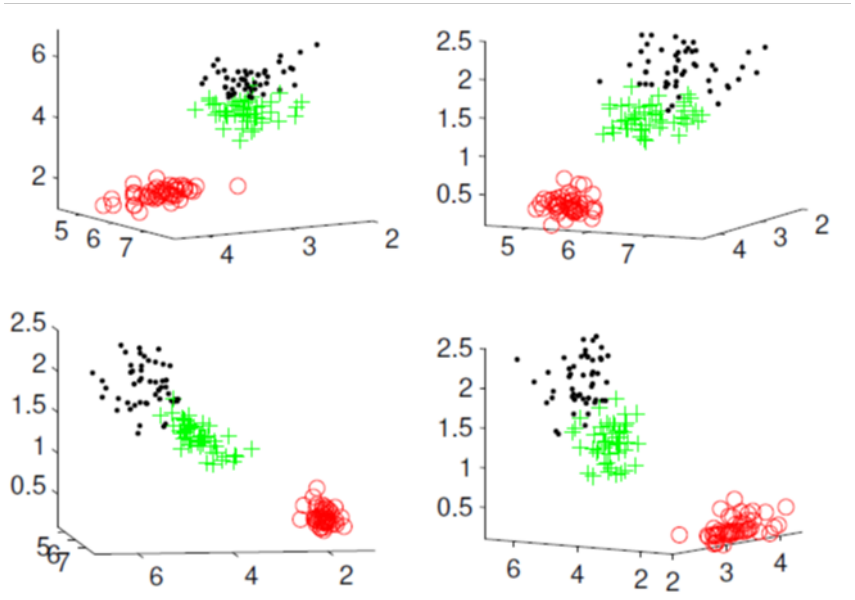


Figure: Three species of the iris data: dims 1, 2 & 3 (top left), dims 1, 2 & 4 (top right), dims 1, 3 & 4 (bottom left) and dims 2, 3 & 4 (bottom right).

How do we interpret these plots? Discuss the different species - different colours.

Pattern Visualization in Data

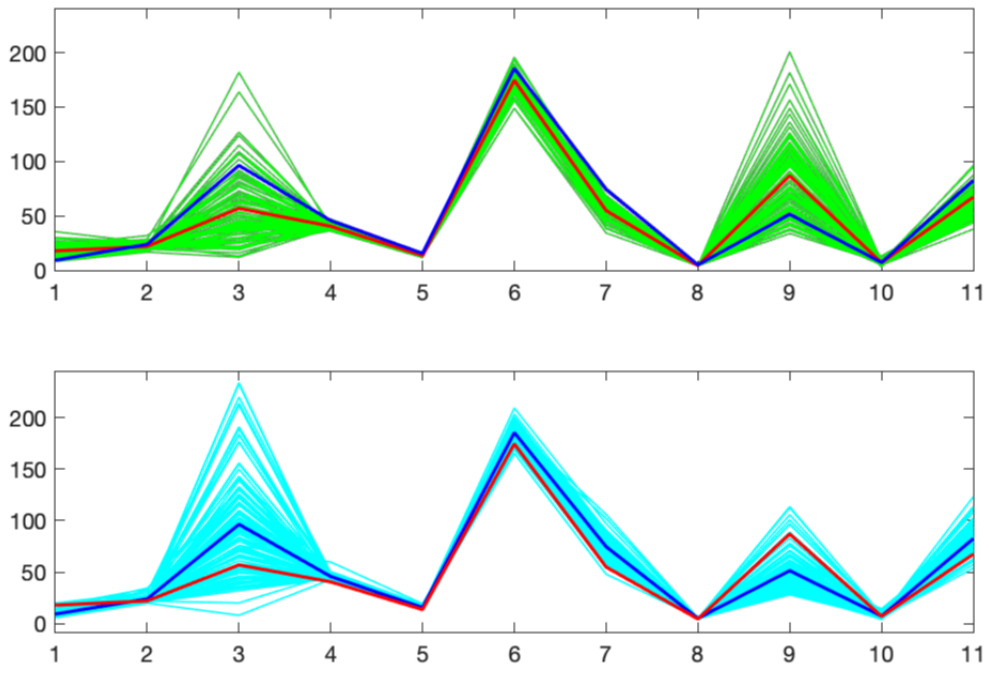


Figure: Athletes data, 100 female athletes (top) and 102 male athletes (bottom) with variables shown on the x-axis, female mean in red, male mean in blue.

What and where are there differences between females and males?

Pattern Visualization in Data

Variables shown on the x-axis, the value of a variable is shown on the y-axis for each observation; boxplots show variability

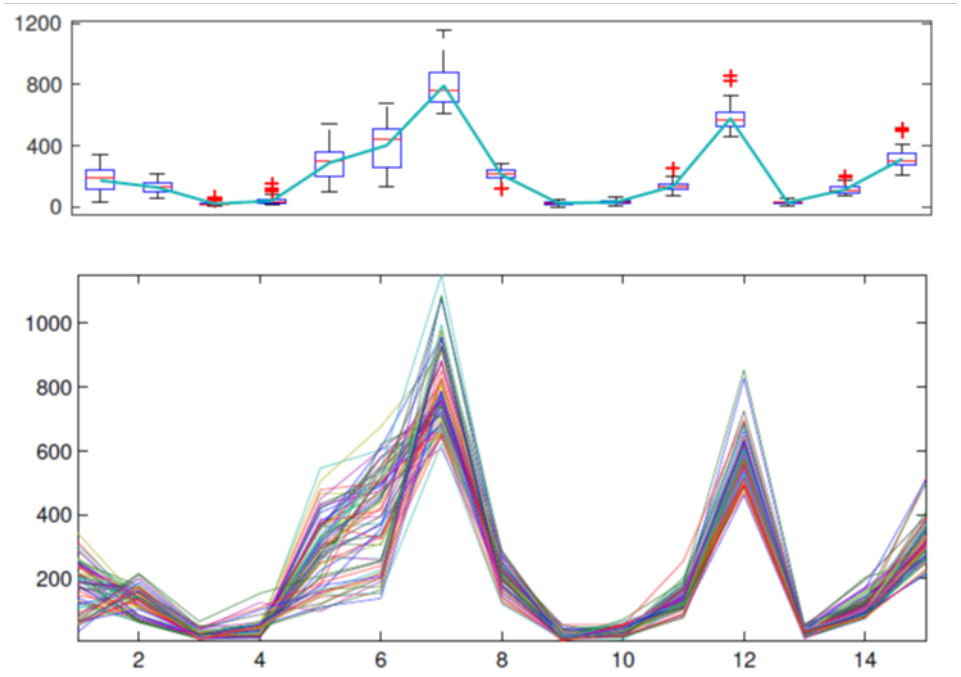


Figure: Parallel coordinate view of the illicit drug market data (bottom) with mean and boxplot for each variable (top).

Pattern Visualization in Data

Visualization of spatial information

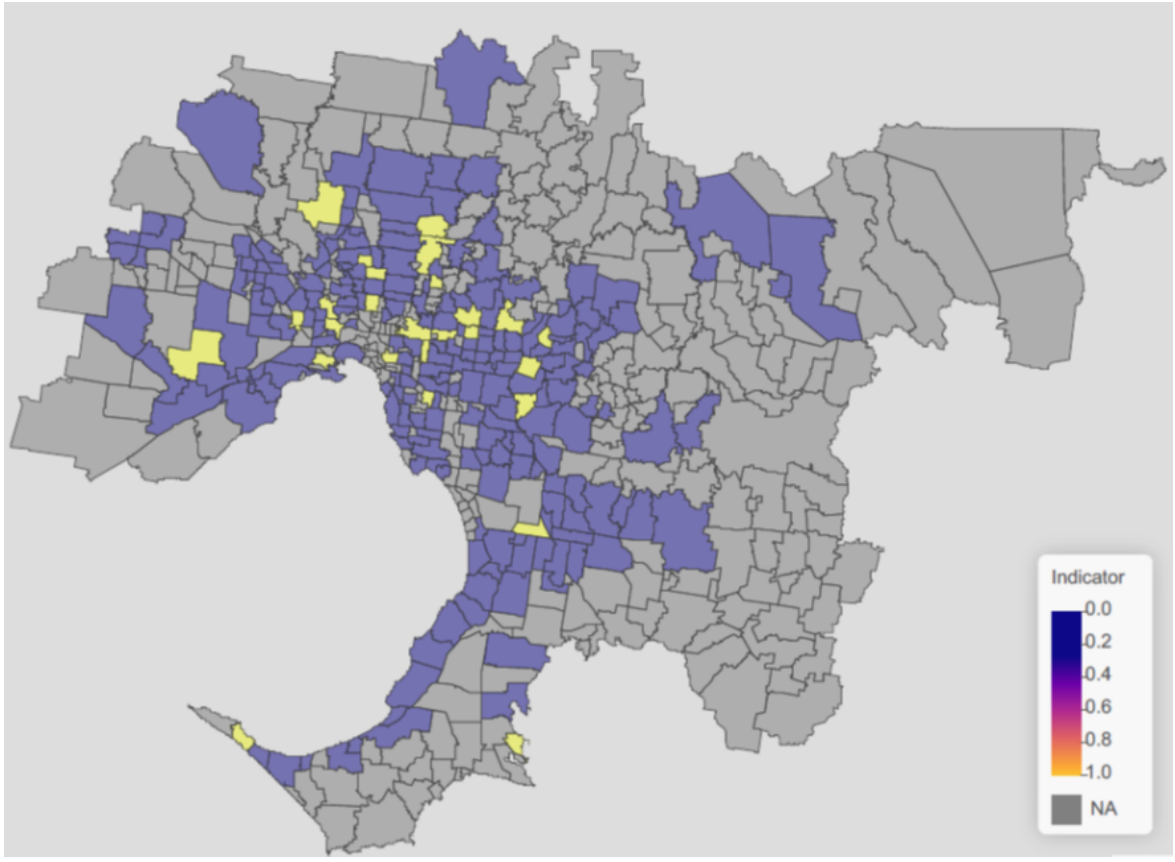


Figure: Relative change of housing price in 2016 compared to 2015 explained by 8 covariates (number of bedrooms, number of bathrooms, land size, building area and built year of the property).

Random Vectors and Data

Some notions

- A single random vector $\mathbf{x}$
- A collection or sample of random vectors $\mathbb{X} \rightarrow$ the random sample or data
- The realized or observed values $\rightarrow$ the observed data

The random vector $\mathbf{x} \in \mathbb{R}^d$ and the $d \times n$ data

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}, \quad \mathbb{X} = [\mathbf{x}_1, ..., \mathbf{x}_n], \text{ and } \quad \mathbf{x}_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{id} \end{bmatrix}, i = 1, ..., n.$$

Expectations and Covariance

For $\mathbf{x} \in \mathbb{R}^d$ the mean $\boldsymbol{\mu}$ is a $d-$dimensional vector and the covariance Σ is a $d \times d$ matrix

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_d \end{bmatrix}, \quad \text{and } \Sigma \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d,1} & \sigma_{d,2} & \cdots & \sigma_d^2 \end{bmatrix},$$

$\sigma_j^2 = \mathrm{var}(x_j)$ is the variance of the j^{th} component of $\mathbf{x}$ and $\sigma_{jk} = \mathrm{cov}(x_j, x_k)$ the covariance between x_j and x_k, $\left(\sigma_{jj} = \sigma_j^2\right)$.
For a random sample $\mathbb{X}$, the sample mean is $\bar{\mathbf{x}}$ and the sample covariance matrix is S

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i, \text{ and } S = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$$

Measure of Association

- $\sigma_{jk} = \mathrm{cov}(x_j, x_k)$ the covariance between x_j and x_k is also related to the Pearson correlation via

$$\rho_{x_j, x_k} = \frac{\mathrm{cov}(x_j, x_k)}{\sigma_j \sigma_k}$$

- It measures the strength of linear association and the sample version is defined as

$$\hat{\sigma}_{jk} = \frac{\sum_{i=1}^{n} (x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n} (x_{ji} - \bar{x}_j)^2 \sum_{i=1}^{n} (x_{ki} - \bar{x}_k)^2}}$$

Random Vectors

We write

$$\mathbf{x} \sim N(\boldsymbol{\mu}, \Sigma)$$

if $\mathbf{x}$ is Gaussian/normal and

$$\mathbf{x} \sim (\boldsymbol{\mu}, \Sigma)$$

if $\mathbf{x}$ is not Gaussian, most of the time ! and use

$$\mathbb{X} \sim (\boldsymbol{\mu}, \Sigma) \text{ and } \mathbb{X} \sim \text{Sam}\,(\bar{\mathbf{x}}, S)$$

to indicate the true parameter of the (unknown) data distribution, and the sample parameters on the right. Centering data is often a first step in an analysis

$$\mathbb{X}_{cent} = \mathbb{X} - \bar{\mathbf{x}}\mathbf{1} = [\mathbf{x}_1 - \bar{\mathbf{x}}, ..., \mathbf{x}_n - \bar{\mathbf{x}}]$$

Smoothed Histograms

Histogram give a first impression of the distribution of data; we use them for one or two variable(s),

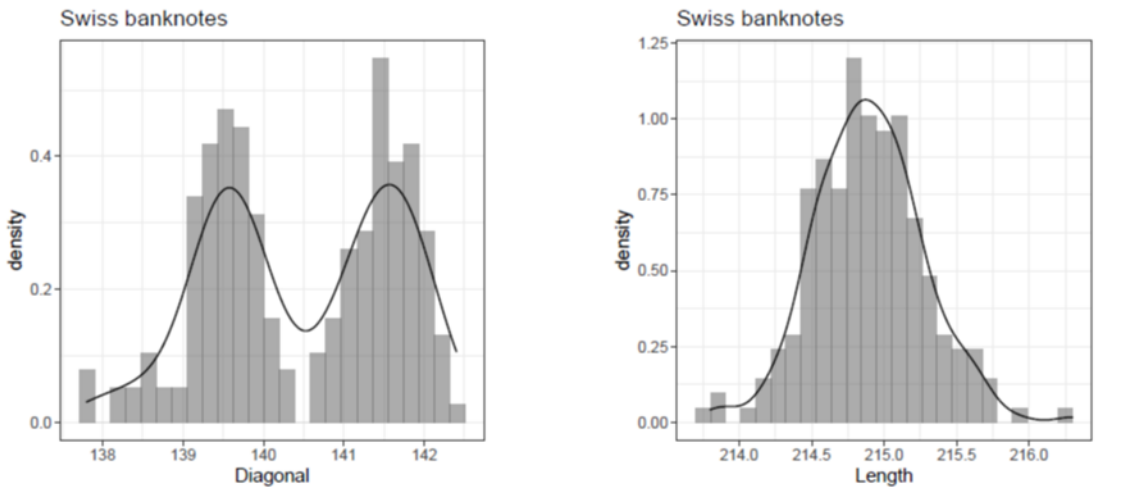


Figure: Smoothed histograms of variables 'diagonal' and 'length' from Swiss bank notes data.

Note the difference in the two distributions: the left is bimodal, the right on is almost symmetric.

Smoothed Histograms

Consider both ‘diagonal’ and ‘length’ of the Swiss bank notes in smoothed histogram or perspective plot

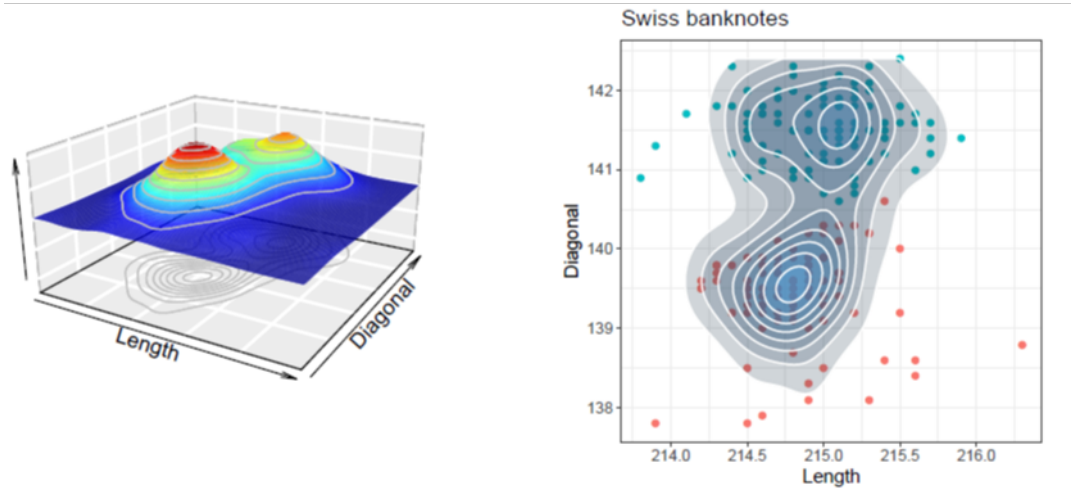


Figure: Perspective and contour plots of variables ‘diagonal’ and ‘length’ from Swiss bank notes data.

How would the 2D plots change if both variables were bimodal?

Smoothed Histograms

Interpretation of histograms:

- Study the shape.
- Compute descriptive statistics.
- Compare the histogram to a standard distribution.

Use this when you want to show how spread out the data are. It is helpful to know when your data are so spread out that it affects the mean.

Spectral Decomposition of the Covariance Matrix

- The covariance Σ can be decomposed into the product

$$\text{pop: } \Sigma = \Gamma\Lambda\Gamma^T \quad \text{data: } S = \widehat{\Gamma}\widehat{\Lambda}\widehat{\Gamma}^T$$

- Λ is a diagonal matrix of eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_d \geq 0$
- $\Gamma = [\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_d]$ is an orthonormal matrix of eigenvectors
- $\|\boldsymbol{\eta}_k\|_2 = 1$ and $\Sigma\boldsymbol{\eta}_k = \lambda_k\boldsymbol{\eta}_k$ for $k = 1, ..., d$

Spectral Decomposition of the Covariance Matrix

Consider an 11×11 covariance matrix with eigenvalues

$$4.99; 2.56; 1.16; 0.89; 0.80; 0.43; 0.11; 0.04; 0.02; 0.005; 0.001$$

and simulate $\mathbb{X} \sim (0, \Sigma)$

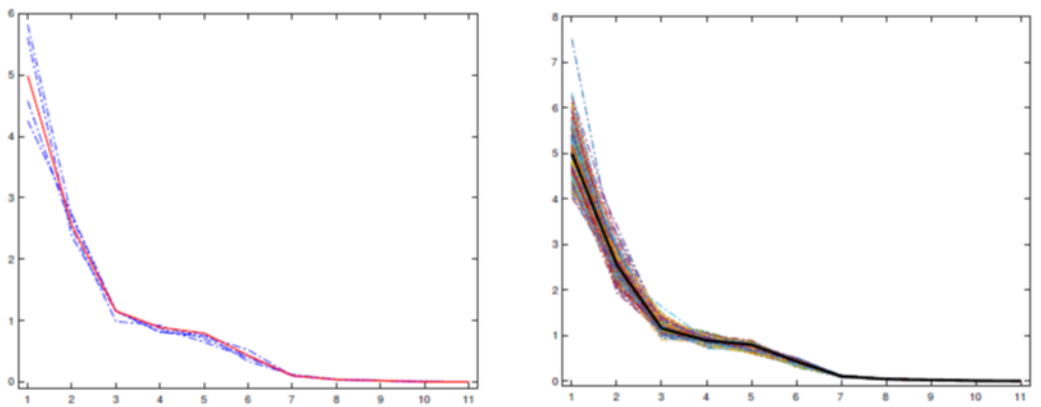


Figure: Eigenvalues of four (left) and 100 (right) sets of simulated data with original eigenvalues shown as solid red (left) and black (right) line.

Working with Spectral Decomposition

- The powers of Σ can easily be obtained from its spectral decomposition $\Sigma^k = \Gamma \Lambda^k \Gamma^T$
- for $k = -1$ and $k = -1/2$, Σ^k is $\Sigma^{-1} = \Gamma \Lambda^{-1} \Gamma^T$ and $\Sigma^{-1/2} = \Gamma \Lambda^{-1/2} \Gamma^T$
- These quantities and their sample versions S^{-1} and $S^{-1/2}$ are widely used
- Λ^{-1} is easy to compute even when Σ^{-1} is not

Simulation

The aim of simulation is to

- generate data randomly from a relevant model;
- apply the method whose performance we are evaluating to the generated data; and
- repeat the above a number of times and analyse the results.

Let's look at the first and third item:

Why simulation helps ? when we apply a method to data

- we do not know the true parameters or the appropriateness of the model; but
- if we simulate the data and apply the analysis method to the data we
  - can estimate the parameters and
  - find out how well our analysis has done in uncovering the true structure or model.

Simulation

Steps taken in a simulation

- choose a reasonable random number generator in R, Python, etc;
- do not automatically use the Gaussian model; but
- choose and describe the mathematical model that is appropriate for your purpose;
- work out how many simulations you need and for what sample sizes - later also for training and testing;
- generate the data from the chosen model;
- carry out the statistical analysis on your generated data; and
- examine the results and evaluate the performance of the simulation and interpret the results.

Gaussian Model for 3D Simulation

Data are generated from three normal distributions - corresponding to three classes, with means $\boldsymbol{\mu}_i$, covariance matrices Σ_i, i=1,2,3

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 1 \\ 0.5 \\ 0 \end{bmatrix}, \quad \boldsymbol{\mu}_3 = \begin{bmatrix} -1 \\ 0.5 \\ 1 \end{bmatrix}$$

and

$$\Sigma_1 \begin{bmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{8} & 0 \\ 0 & 0 & \frac{1}{8} \end{bmatrix}, \quad \Sigma_2 \begin{bmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} \end{bmatrix}, \quad \Sigma_1 \begin{bmatrix} \frac{1}{8} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} \end{bmatrix}.$$

Result of Simulation from Gaussian Model

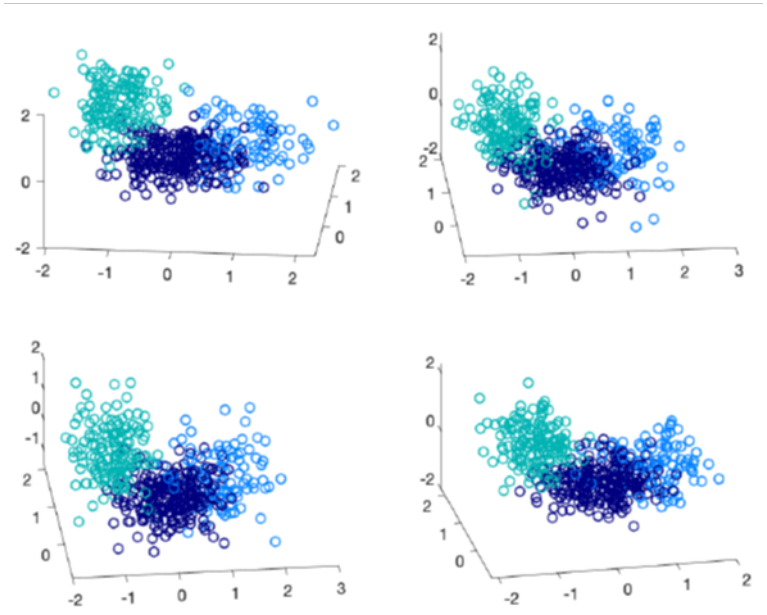


Figure: Simulated data in 3D based on 250, 100 and 150 random vectors from class 1, 2, 3 resp.

Of special interest is the sampling variability visible in the 4 simulations.

Simulation

To share data and results with others and re-use previous simulations and data analyis, the simulations themselves should be reproducible.

- to make them reproducible, use seeds - be aware: this may not be comparable between different generators or different software;
- this will allow you to repeat simulation or re-use and compare with new simulations;
- these ideas also apply to drawing a subsample from a data set; and special attention is required to sampling with or without replacement - why?

Resampling

Resampling is used to select a new sample from an orginal sample

- sampling with replacement is one of the key ideas of the bootstrap which is used, e.g. in Random Forests;
- 'oversampling' may be required to augment a data set with extra observations, e.g. in classification when one or more classes are rare - why?

Performance Evaluation and Optimism

Aims in statistical modelling and data analysis include:

- describing the data; and
- using the description to predict unseen values in similar circumstances.

In classification 'describing' includes
finding a rule

- which is good at separating the classes; and
- which is good at predicting which class a new individual belongs to.

Classification and Performance Evaluation

The two problems are very different and relate to

- classification for the existing data and to
- prediction for unseen observations from the same classes.

Different performance measures and result in different answers and there is not usually one method or measure that is best.

Performance Evaluation

- Might expect that a classification rule that behaves well on observed data will also behave well on new data of the same type;
- This refers to an optimism built into the process of estimating the accuracy of a rule on the same data as the rule was developed from.
- If we use the rule on a similar but different data, the rule will not have been optimised for the new data set and we must therefore expect a possibly worse result.

We need to examine: how accurate do we expect the rule to be on new data?

Optimism and Simulation

- Assume we have chose a classification rule and a measure of accuracy which has a misclassification rate of 8.4%.
- We test the rule on 1000 simulations from the same classes and same numbers an record the misclassification rate in the form of a density plot.

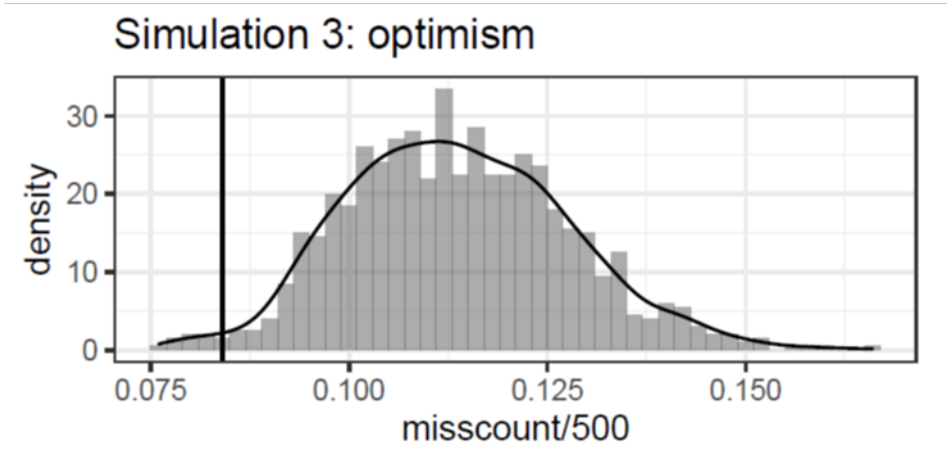


Figure: Classification error for the third simulated data set in 3D and with 3 classes.