Applied Statistics - Notes - v0.1.0

260236

March 2025

Preface

Every theory section in these notes has been taken from two sources:

- The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. [2]
- An Introduction to Statistical Learning: with Applications in Python. [3]
- Applied Multivariate Statistical Analysis. [4]
- Course slides. [5]

About:

GitHub repository

These notes are an unofficial resource and shouldn't replace the course material or any other book on applied statistics. It is not made for commercial purposes. I've made the following notes to help me improve my knowledge and maybe it can be helpful for everyone.

As I have highlighted, a student should choose the teacher's material or a book on the topic. These notes can only be a helpful material.

Contents

1 Business Data Analytics

1.1 Multivariate Descriptive Statistics

When we move from **analyzing** a single variable (univariate analysis) to **multiple variables at once**, we enter the realm of **Multivariate (MV) analysis**. A natural question arises: *Is multivariate analysis just a replication of univariate analysis across several variables?*

The answer is no, multivariate analysis introduces new and fundamental questions that cannot be answered by simply analyzing variables individually. The **core focus** shifts to **understanding how these variables interact with each other**. Specifically, we are concerned with the **dependence** and **correlation between variables**.

Covariance: Measuring Joint Variability

To capture how two variables vary together, we use **Covariance**. The **Sample Covariance** between variables x_j and x_k is calculated as:

$$\mathrm{cov}_{jk} = s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \tag{1}$$

- $s_{jk} = 0 \Rightarrow$ implies that there is **no linear relationship** between the two variables.
- $s_{jk} > 0 \Rightarrow$ suggests that **as one variable increases**, the **other tends to increase**.
- $s_{jk} < 0 \Rightarrow$ one **variable** tends to **decrease when the other increases**.

Covariance Is Not Standardized

The **value of covariance is not standardized**, it **depends on the units of measurement**, which makes comparisons difficult. For example:

- Suppose we're measuring
  - Height in centimeters
  - Weight in kilograms
- The covariance between height and weight will be expressed in *centimeter-kilogram* units.

Now imagine we convert height to meters. The covariance value changes, because now you're multiplying meters × kilograms instead of centimeters × kilograms. Even though the **relationship between height and weight hasn't changed**, the **numerical value of covariance does change due to this unit change**. **Because of unit dependency, it's hard to compare covariances between different variable pairs**. Finally, it is **hard to interpret the magnitude of covariance in any absolute sense** (e.g., is 50 a large covariance or small? It depends on the units!).

Correlation: Standardized Covariance

To standardize covariance and **measure the strength of a linear relationship** on a scale between -1 and 1, we use the **Correlation** coefficient, defined as:

$$\mathrm{cor}_{jk} = r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}} = \frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}} \tag{2}$$

This formula divides the covariance by the product of the standard deviations of the two variables, giving a **unitless value**:

- $r = 0$: No linear correlation
- $r > 0$: Positive correlation (**both variables increase or decrease together**)
- $r < 0$: Negative correlation (**one increases while the other decreases**)
- $|r| = 1$: Perfect correlation (**exact linear dependence**)

Thus, correlation not only **reveals the direction of the relationship** but also its **strength**.

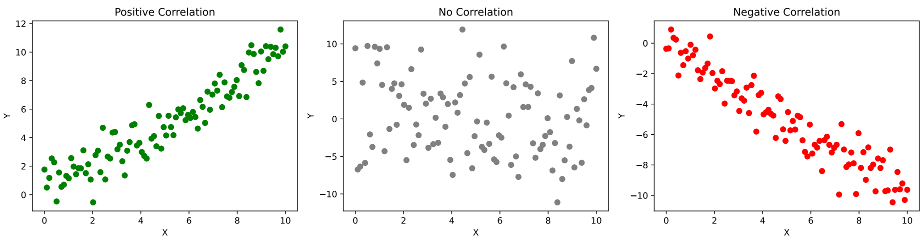

Figure 1: Direction of correlation: positive (left), none (center), negative (right).

Describing MV Data: Vector and Matrices

When analyzing multivariate data:

- We compute the **vector of Sample Means**:

$$\bar{\mathbf{X}} = [\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_p] \tag{3}$$

- And the **Variance-Covariance matrix S**, which **summarizes the covariances between all pairs of variables**:

$$S = \begin{bmatrix} s_{11} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_{pp} \end{bmatrix} \tag{4}$$

- Alternatively, we can use the **Correlation matrix R**, where **all diagonal elements are 1** (because each variable is perfectly correlated with itself) and **off-diagonal elements are correlation coefficients**:

$$R = \begin{bmatrix} 1 & \cdots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{p1} & \cdots & 1 \end{bmatrix} \tag{5}$$

Scatterplots - Visualizing Variable Pairs

One of the most intuitive and widely used tools in multivariate analysis is the **2D Scatterplot**. Each plot shows how two variables relate to each other:

- ✔ Clusters or linear trends can indicate correlation or dependence.
- ✔ Scatterplots are **ideal for spotting positive, negative, or no correlation** visually.

However, scatterplots have a **limitation**: they **only** allow us to **analyze two variables at a time**. When dealing with many variables, the **number of possible pairings becomes large**, making it **difficult to read or interpret** the scatterplots individually. This is where quantitative measures (like correlation matrices) and higher-dimensional graphics come into play.

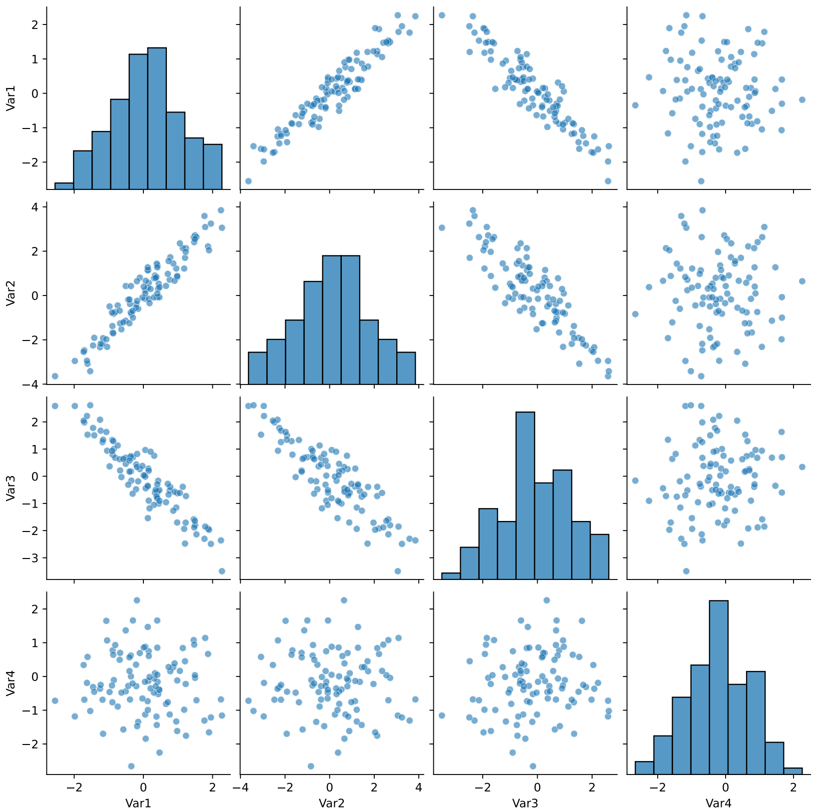

Figure 2: Scatterplot matrix of four variables. This scatterplot matrix displays all pairwise relationships among four variables:

- Diagonal plots (top-left to bottom-right): Histograms showing the distribution of each variable.
- Off-diagonal plots: 2D scatterplots illustrating the relationship between each pair of variables.

Rotated Plots in 3D - Capturing Complexity

When dealing with **three variables**, we can **extend scatterplots into 3D space** using **Rotated plots**. These visualizations allow us to:

- ✔ **Explore interdependencies** among three variables at once.
- ✔ **Gain spatial insight** into how data points spread in three-dimensional space.
- ✔ Observe **complex patterns that are invisible in 2D**.

Yet again, as we move **beyond three variables**, visualizing becomes **impractical**, our brains cannot easily comprehend 4D or higher dimensions. Hence, dimensionality reduction techniques like PCA are often used alongside visualizations to make high-dimensional data “digestible”.

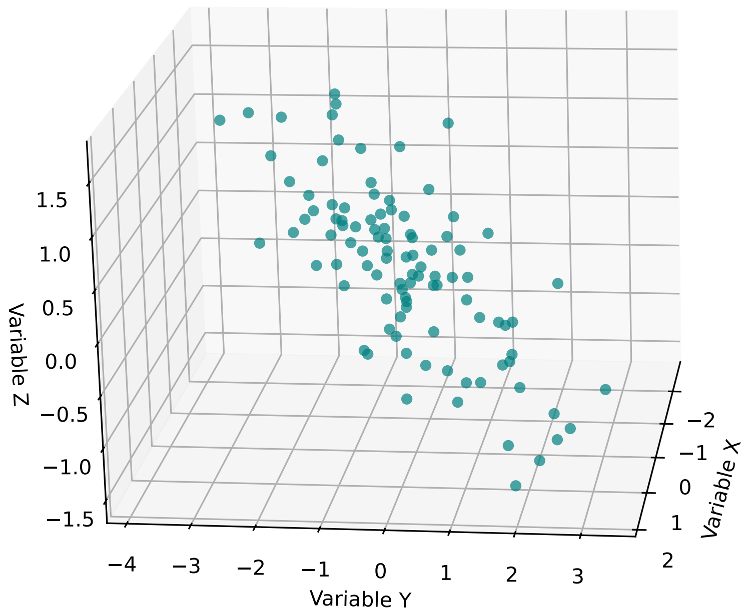

Figure 3: A simple rotated plots in 3D.

Star Plots - Shape-Based Comparison

Star plots offer a creative way to **represent multivariate data**:

- **Each variable is represented as a ray** (spoke) starting from a central point.
- The **length of each ray** corresponds to the **value** of that variable.
- When **rays are connected**, they form a "star-like shape" **unique** to each observation.

This method is **excellent for comparing patterns** between observations:

✔ **Similar shapes** suggest **similar data profiles**.

✔ **Differences** in shape can **quickly highlight outliers** or clusters.

However, star plots have limitations:

✘ They **do not quantify correlation**.

✘ The **direction** and **magnitude** of relationships between variables are **not explicit**.

✘ They are better for **visual pattern recognition** than for precise statistical analysis.

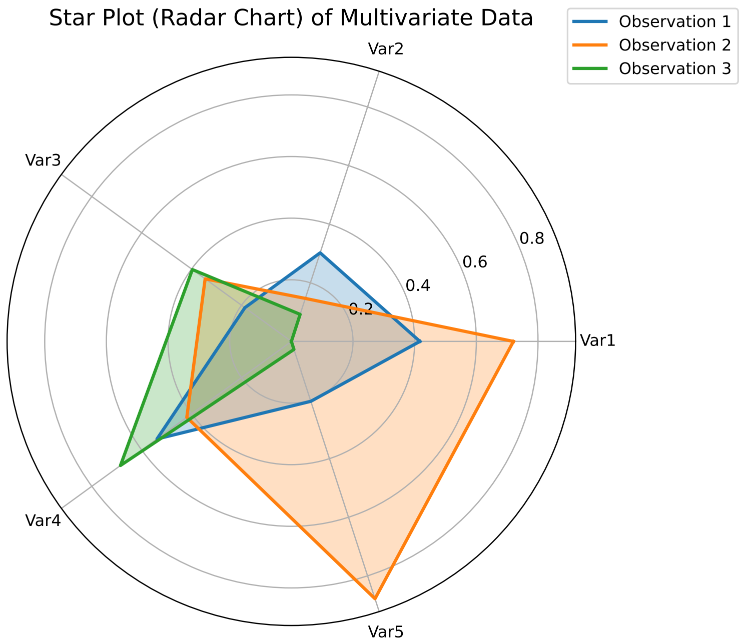

Figure 4: Star Plot (Radar Chart) of Multivariate Data.

Chernoff Faces - Human-Centric Visualization

Chernoff faces [1] are an **innovative visualization method** where **multivariate data is represented as a human face**:

- Each **variable controls a facial feature** (e.g., mouth curvature, eye size, nose length).
- **People are naturally attuned to recognizing faces** and subtle differences in expressions.
- Hence, Chernoff faces **leverage human perception** for quickly comparing **multivariate observations**.

Despite being engaging, Chernoff faces also have **drawbacks**:

✗ They **do not provide numerical precision**.

✗ The **mapping of variables to facial features** can be **arbitrary**.

✗ They work best as a **qualitative summary tool** rather than for deep statistical inference.

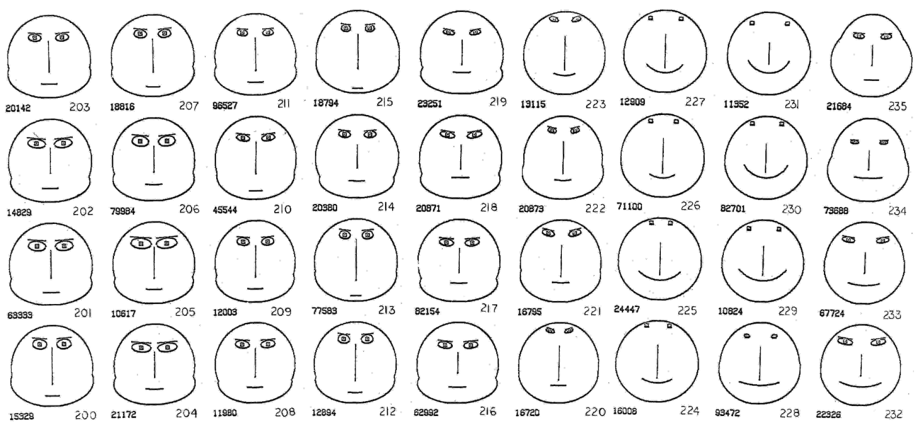

Figure 5: Some Chernoff faces. [1]

Graphic Type	Strengths	Limitations
Scatterplots	Clear view of pairwise relationships	Hard to scale beyond 2 variables
3D Plots	Visualizes 3-variable interaction	Limited to 3 dimensions, requires rotation
Star Plots	Quick shape-based comparison across variables	No quantification, poor at showing correlations
Chernoff Faces	Leverages facial perception for comparison	Subjective, lacks precision

Table 1: When and why to use graphics.

1.2 Dimensionality Reduction

The Challenge: Data in High Dimensions

In many real-world problems, we **collect multiple variables for each observation**. For example, in a medical study, a patient might be described by age, weight, blood pressure, and dozens of test results. This leads to **high-dimensional data**, where **each observation is a point in a complex, multi-dimensional space** (formally, a Euclidean space of dimension p).

The problem? As the **number of variables** (p) **increases**, the **data becomes harder to visualize, interpret, and model**:

- Some **variables** might be **redundant** or **highly correlated**.
- **Computations** become more **expensive**.
- **Patterns become obscured** by the complexity.

Goal of Dimensionality Reduction

We want to **summarize the data using fewer variables**, say k derived variables (with $k < p$), that still **retain most of the information**. This process is a balancing act:

- **Clarity**: fewer variables make data easier to understand and visualize.
- **Risk of oversimplification**: reducing dimensions too much can cause loss of important information.

The key concept here is that in statistics, **information is variability**. The more variability we retain from the original data, the more information we preserve.

Example 1: Blood Cells

Imagine we measure thickness and diameter for a set of red blood cells:

- Each cell = one observation with two variables.
- We can represent this as a table (numbers) or as a 2D scatterplot.

Now we ask ourselves: *Can we describe these cells using only one feature instead of two?*
If we choose only diameter or only thickness, we'll lose detail:

- Some cells will appear more similar than they really are.
- We miss variability that distinguishes them.

So, we seek a better single feature, one that captures the most variation possible from both thickness and diameter combined.

The Statistical Insight: Maximize Variability

Rather than randomly picking a feature, we **analyze the directions along which the data varies the most**.

1. First, we find the **direction of maximum spread**.
2. Then, the **second most spread direction**, orthogonal to the first.

This is the essence of **Principal Component Analysis (PCA)**, a **dimensionality reduction technique that finds the best directions** (linear combinations of variables) **to project the data**, while **maximizing retained variability**.

Dimensionality Reduction in Practice

Let's formalize the idea:

- We start with a **data matrix** X of shape $n \times p$ (n observations, p variables).
- The goal is to **obtain a new matrix** M of shape $n \times k$, with $k < p$, that **captures most of the variability**.
- The **difference** between X and M is **residual variation**, the **information lost**.

In summary, **Dimensionality Reduction** is about simplifying complexity: **transforming a large set of variables into a smaller**, more interpretable set **without losing the essence of the data**. It's central to data exploration, preprocessing, and modeling, especially when working with high-dimensional datasets.

References

[1] Herman Chernoff. The use of faces to represent points in k-dimensional space graphically. *Journal of the American statistical Association*, 68(342):361–368, 1973.

[2] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York, 2009.

[3] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in Python*. Springer Texts in Statistics. Springer New York, 2013.

[4] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Applied Multivariate Statistical Analysis. Pearson Prentice Hall, 2007.

[5] Beraha Mario. Applied statistics. Slides from the HPC-E master's degree course on Politecnico di Milano, 2024-2025.

Index