Multivariate Statistical Analysis

Math4424
HKUST

Kani Chen (Instructor)

Chapter 1.
Introduction (Aspects of Multivariate Analysis.)

Multivariate analysis generally refers to a range of statistical techniques/methods which primarily involves data with several variables, with the objective of investigating the dependence structure or relations between the involved variables.

In this Chapter, we briefly review some basic numerical/descriptive and graphical methods in exploring univariate or multivariate data.

Example 1.1. Ten American Companies.

Ten (biggest) American companies are listed with three variables sales, profit and asset. The details are shown in the R-program. (See Appendix A.)

□

1. Descriptive statistics.

The descriptive statistics are characterization/caliberation of data through some summary numbers. (e.g., mean, variance, etc.)

Suppose the data is presented as

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

Such a presentation implies, as our convention, there are p variables, n observations. In particular, $(x_{i1}, x_{i2}, ..., x_{ip})$ are the i-th observation of the p variables, and the observations of the k-th variable are $x_{1k}, ..., x_{nk}$. The following are some of the typical and elementary descriptive statistics concerning mean and variances:

Sample mean (for variable k): $\bar{x}_k = (1/n)\sum_{i=1}^n x_{ik}$.

Sample variance (for variable k): $s_k^2 = \frac{1}{n-1}\sum_{i=1}^n (x_{ik} - \bar{x}_k)^2$.

Sample standard deviation (for variable k): $s_k = \sqrt{s_k^2}$.

Sample covariances (for variables k and l): $s_{kl} = \frac{1}{n-1}\sum_{i=1}^n (x_{ik} - \bar{x}_k)(x_{il} - \bar{x}_l)$.

Sample correlation (for variables k and l): $r_{kl} = \frac{s_{kl}}{s_k s_l} = \frac{s_{kl}}{\sqrt{s_{kk} s_{ll}}}$.

(Caution: $s_{kk} = s_k^2$).

The matrix presentations are

$$\bar{X} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \bar{x}_p \end{pmatrix} \quad \mathbf{S} = \begin{pmatrix} s_{11} & \cdots & s_{1p} \\ \vdots & \vdots & \vdots \\ s_{p1} & \cdots & s_{pp} \end{pmatrix}_{p\times p} \quad \mathbf{R} = \begin{pmatrix} r_{11} & \cdots & r_{1p} \\ \vdots & \vdots & \vdots \\ r_{p1} & \cdots & r_{pp} \end{pmatrix}_{p\times p}.$$

2. Graphical methods.

To investigate the distribution of one variable, one can use histogram, boxplot, density plot, qqnorm, etc.

To investigate the distribution of two variables, particularly caring of their relations, one can use scatter plot, qqplot, etc.

When we face more than two variables, as in our course, graphical tools of higher dimension are not readily available. Still, we may draw one variable or two at a time and use graphic techniques as described above to investigate their marginal distributions of any one or any pair of variables.

We introduce two approaches, Chernoff faces and star plots, that are designed to check the characteristics of the *observations.* (see Appendix A.)

Remark. (regarding qqplot and qqnorm). When use qqplot(x, y) to plot the quantiles of y against those of x, we are investigating whether the distribution of y and distribution of x has certain connection. If y has the same distribution of a linear transformation of x, say $ax + b$, then the plot is close to a straight line with intercept b and slope a.

In particular, qqnorm(x) is a plot of the quantiles of x against the corresponding quantiles of standard normal distribution $N(0, 1)$. Therefore, if the plot is close to a straight line with intercept b and slope a, that indicates x is approximately normal with mean b and variance a^2.

The following is a (sloppy) mathematical proposition in support of the above justification of qqplot. Note that $F^{-1}(t)$ is the quantile of distribution F.

Suppose $X \sim F_X(\cdot)$ and $Y \sim F_Y(\cdot)$. Then $aX + b$ has the same distribution as Y for some $a > 0$ if and only if $F_Y^{-1}(t) = aF_X^{-1}(t) + b$ for all $t \in [0, 1]$.

Proof. $\Longrightarrow$. Given any s,

$$F_Y(s) = P(Y \le s) = P(aX + b \le s) = P(X \le (s - b)/a) = F_X((s - b)/a).$$

Then, $F_X^{-1}(F_Y(s)) = (s - b)/a$ and, as a result, $F_X^{-1}(t) = (1/a)(F_Y^{-1}(t) - b)$.

$\Longleftarrow$: DIY. □

3. Statistical distance.

Imagine that we can plot n observations of p variables. They are n points in R^p, the real space of p-dimension. For any point in the space, we wish to use certain distance to measure how the point fit into the pattern of the data. Let first recall the Euclidean distances.

3.1. Euclidean distance: For two points X and Y in R^p, the Euclidean norms are $\|X\| = \sqrt{\sum_{i=1}^p x_i^2}$ and $\|Y\| = \sqrt{\sum_{i=1}^p y_i^2}$, where x_i and y_i are the cooridinates of X and Y, respectively. Note that the norm of a point actually measures the Euclidean distrance between the point and the origin of the space. The Euclidean distance between X and Y is

$$\|X - Y\| = \sqrt{\sum_{i=1}^p (x_i - y_i)^2}.$$

3.2. Statistical distance. Recall that the data are presented as

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

The "center" of the data are here represented by the sample mean

$$\bar{X} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \bar{x}_p \end{pmatrix}$$

Consider an arbitrary point Y (not random variable and not necessarily one of the observations) in R^p, say,

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}.$$

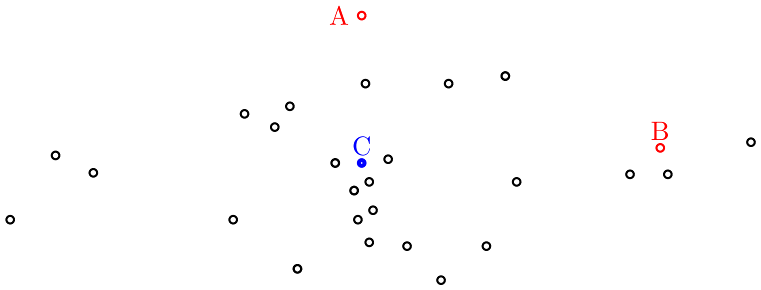
The "statistical norm" of Y is defined as

$$\|Y\| = \sqrt{(y_1 - \bar{x}_1 \ \cdots \ y_p - \bar{x}_p)\mathbf{S}^{-1}\begin{pmatrix} y_1 - \bar{x}_1 \\ \vdots \\ y_p - \bar{x}_p \end{pmatrix}} = \sqrt{(Y - \bar{X})'\mathbf{S}^{-1}(Y - \bar{X})}.$$

And the statistical distances between two points Y and $\tilde{Y}$ is $\sqrt{(Y - \tilde{Y})'\mathbf{S}^{-1}(Y - \tilde{Y})}$.

Here the center of the universe (R^p) is taken to be $\bar{X}$, playing the role of the origin of R^p. The universe is affine transformed by $\mathbf{S}$, so that the length along the directions with diverse observations are contracted and that along the directions with concentrated observations are extended. In other words, the more elastic directions are discounted in computing distances or norms.

The smaller the statistical norm of a point or an observation, the better the point or the observation fits into the pattern of the data.



Compare point A and point B in the above figure. One can easily see that A is closer to C, the "center", than B is, in terms of Euclidean distance. In terms of statistical distance, B is closer to the center. Indeed, we may visually judge that B fits the data better than A. In fact, A may be regarded as an outlier, but B should not.

We shall skip Chapters 2 and 3 in the textbook.