Chapter 1

Introduction

1.1 Introduction

Multivariate analysis is a set of statistical techniques used to analyze correlated data containing observations on $p \geq 2$ random variables measured on a set of n cases. Let $\boldsymbol{x} = (x_1, ..., x_p)^T$ where $x_1, ..., x_p$ are p random variables. Usually context will be used to decide whether $\boldsymbol{x}$ is a random vector or the observed random vector. For multivariate location and dispersion the ith case is $\boldsymbol{x}_i = (x_{i,1}, ..., x_{i,p})^T$.

Notation: Typically lower case boldface letters such as $\boldsymbol{x}$ denote column vectors while upper case boldface letters such as $\boldsymbol{S}$ denote matrices with 2 or more columns. An exception may occur for random vectors which are usually denoted by $\boldsymbol{x}$, $\boldsymbol{y}$ or $\boldsymbol{z}$. If context is not enough to determine whether $\boldsymbol{x}$ is a random vector or an observed random vector, then $\boldsymbol{X} = (X_1, ..., X_p)^T$ and $\boldsymbol{Y}$ will be used for the random vectors, and $\boldsymbol{x} = (x_1, ..., x_p)^T$ for observed value of the random vector. This notation is used in Chapter 3 in order to study the conditional distribution of $\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}$. An upper case letter such as Y will usually be a random variable. A lower case letter such as x_1 will also often be a random variable. An exception to this notation is the generic multivariate location and dispersion estimator $(T, \boldsymbol{C})$ where the location estimator T is a $p \times 1$ vector such as $T = \overline{\boldsymbol{x}}$. $\boldsymbol{C}$ is a $p \times p$ dispersion estimator and conforms to the above notation. Another exception is in Chapter 3 where

Assume that the data $\boldsymbol{x}_i$ has been observed and stored in an $n \times p$ matrix

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_p \end{bmatrix}$$

where the ith row of $\boldsymbol{W}$ is the ith case $\boldsymbol{x}_i^T$ and the jth column $\boldsymbol{v}_j$ of $\boldsymbol{W}$ corresponds to n measurements of the jth random variable x_j for $j = 1, ..., p$.

Often the n rows corresponding to the n cases are assumed to be iid or a random sample from some multivariate distribution. The p columns correspond to n measurements on the p correlated random variables $x_1, ..., x_p$. The n cases are $p \times 1$ vectors while the p columns are $n \times 1$ vectors.

Methods involving one response variable will not be covered in depth in this text. Such models include multiple linear regression, many experimental design models and generalized linear models. Discrete multivariate analysis = categorical data analysis will also not be covered.

Most of the multivariate techniques studied in this book will use estimators of multivariate location and dispersion. Typically the data will be assumed to come from a continuous distribution with a joint probability distribution function (pdf). Multivariate techniques that examine correlations among the p random variables $x_1, ..., x_p$ include principal component analysis, canonical correlation analysis and factor analysis. Multivariate techniques that compare the n cases $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ include discriminant analysis and cluster analysis. *Data reduction* attempts to simplify the multivariate data without losing important information. Since the data matrix $\boldsymbol{W}$ has np terms, *data reduction* is an important technique. Prediction and hypothesis testing are also important techniques. Hypothesis testing is important for multivariate regression, Hotelling's T^2 test, and MANOVA.

Robust multivariate analysis consists of i) techniques that are robust to nonnormality or ii) techniques that are robust to outliers. Techniques that are robust to outliers tend to have some robustness to nonnormality. The classical covariance matrix $\boldsymbol{S}$ is very robust to nonnormality, but is not robust to outliers. Large sample theory is useful for both robust techniques. See Section 3.4.

1.2 Things That Can Go Wrong with a Multivariate Analysis

In multivariate analysis, there is often a training data set used to predict or classify data in a future data set. Many things can go wrong. For classification and prediction, it is usually assumed that the data in the training set is from the same distribution as the data in the future set. Following Hand (2006), this crucial assumption is often not justified.

Population drift is a common reason why the above assumption, which assumes that the various distributions involved do not change over time, is violated. Population drift occurs when the population distribution does change over time. As an example, perhaps pot shards are classified after being sent to a lab for analysis. It is often the case that even if the shards are sent to the same lab twice, the two sets of lab measurements differ greatly. As another example, suppose there are several variables being used to produce greater yield of a crop or a chemical. If one journal paper out of 50 (the training set) finds a set of variables and variable levels that successfully increases yield, then the next 25 papers (the future set) are more likely to use variables and variable levels similar to the one successful paper than variables and variable levels of the 49 papers that did not succeed. Hand (2006) notes that classification rules used to predict whether applicants are likely to default on loans are updated every few months in the banking and credit scoring industries.

A second thing that can go wrong is that the training or future data set is distorted away from the population distribution. This could occur if outliers are present or if one of the data sets is not a random sample from the population. For example, the training data set could be drawn from three hospitals, and the future data set could be drawn from two more hospitals. These two data sets may not represent random samples from the same population of hospitals.

Often problems specific to the multivariate method can occur. Often simpler techniques can outperform sophisticated multivariate techniques because the user of the multivariate method does not have the expertise to get the most out of the sophisticated technique. For supervised classification, Hand (2006) notes that there can be error in class labels, arbitrariness in class definitions and data sets where different optimization criteria lead to very different classification rules. Hand (2006) suggests that simple rules such as linear discriminant analysis may perform almost as well or better than sophisticated classification rules because of all of the possible problems. See Chapter 8.

1.3 Some Matrix Optimization Results

The following results will be useful throughout the text. Let $\boldsymbol{A} > 0$ denote that $\boldsymbol{A}$ is a positive definite matrix.

Theorem 1.1. Let $\boldsymbol{B} > 0$ be a $p \times p$ symmetric matrix with eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{e}_1), ..., (\lambda_p, \boldsymbol{e}_p)$ where $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p > 0$ and the orthonormal eigenvectors satisfy $\boldsymbol{e}_i^T \boldsymbol{e}_i = 1$ while $\boldsymbol{e}_i^T \boldsymbol{e}_j = 0$ for $i \neq j$. Let $\boldsymbol{d}$ be a given $p \times 1$ vector and let $\boldsymbol{a}$ be an arbitrary nonzero $p \times 1$ vector. See Johnson and Wichern (1988, p. 64-65, 184).

a) $\max_{\boldsymbol{a} \neq \boldsymbol{0}} \dfrac{\boldsymbol{a}^T \boldsymbol{d}\boldsymbol{d}^T \boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{B} \boldsymbol{a}} = \boldsymbol{d}^T \boldsymbol{B}^{-1} \boldsymbol{d}$ where the max is attained for $\boldsymbol{a} = c\boldsymbol{B}^{-1}\boldsymbol{d}$ for any constant $c \neq 0$. Note that the numerator $= (\boldsymbol{a}^T \boldsymbol{d})^2$.

b) $\max_{\boldsymbol{a} \neq \boldsymbol{0}} \dfrac{\boldsymbol{a}^T \boldsymbol{B} \boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{a}} = \max_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^T \boldsymbol{B} \boldsymbol{a} = \lambda_1$ where the max is attained for $\boldsymbol{a} = \boldsymbol{e}_1$.

c) $\min_{\boldsymbol{a} \neq \boldsymbol{0}} \dfrac{\boldsymbol{a}^T \boldsymbol{B} \boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{a}} = \min_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^T \boldsymbol{B} \boldsymbol{a} = \lambda_p$ where the min is attained for $\boldsymbol{a} = \boldsymbol{e}_p$.

d) $\max_{\boldsymbol{a} \perp \boldsymbol{e}_1, ..., \boldsymbol{e}_k} \dfrac{\boldsymbol{a}^T \boldsymbol{B} \boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{a}} = \max_{\|\boldsymbol{a}\|=1, \boldsymbol{a} \perp \boldsymbol{e}_1, ..., \boldsymbol{e}_k} \boldsymbol{a}^T \boldsymbol{B} \boldsymbol{a} = \lambda_{k+1}$ where the max is attained for $\boldsymbol{a} = \boldsymbol{e}_{k+1}$ for $k = 1, 2, ..., p-1$.

e) Let $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ be the observed sample mean and sample covariance matrix where $\boldsymbol{S} > 0$. Then $\max_{\boldsymbol{a} \neq \boldsymbol{0}} \dfrac{\boldsymbol{a}^T (\overline{\boldsymbol{x}} - \boldsymbol{\mu})(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{S} \boldsymbol{a}} = n(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}) = T^2$ where the max is attained for $\boldsymbol{a} = c\boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu})$ for constant $c \neq 0$.

f) Let $\boldsymbol{A}$ be a $p \times p$ symmetric matrix. Then $\max_{\boldsymbol{a} \neq \boldsymbol{0}} \dfrac{\boldsymbol{a}^T \boldsymbol{A} \boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{B} \boldsymbol{a}} = \lambda_1(\boldsymbol{B}^{-1}\boldsymbol{A})$, the largest eigenvalue of $\boldsymbol{B}^{-1}\boldsymbol{A}$.

1.4 The Location Model

The *location model*

$$Y_i = \mu + e_i, \quad i = 1, \ldots, n \tag{1.1}$$

is a special case of the multivariate location and dispersion model with $p = 1$. The location model is often summarized by obtaining point estimates and confidence intervals for a location parameter and a scale parameter. Assume that there is a sample $Y_1, \ldots, Y_n$ of size n where the Y_i are iid from a distribution with median MED(Y), mean $E(Y)$, and variance $V(Y)$ if they exist. Also assume that the Y_i have a cumulative distribution function (cdf) F that is known up to a few parameters. For example, Y_i could be normal, exponential, or double exponential. The location parameter μ is often the population mean or median while the scale parameter is often the population standard deviation $\sqrt{V(Y)}$. The ith *case* is Y_i.

Point estimation is one of the oldest problems in statistics and four of the most important statistics for the location model are the sample mean, median, variance, and the median absolute deviation (mad). Let $Y_1, \ldots, Y_n$ be the random sample; ie, assume that $Y_1, ..., Y_n$ are iid.

Definition 1.1. The *sample mean*

$$\overline{Y} = \frac{\sum_{i=1}^n Y_i}{n}. \tag{1.2}$$

The sample mean is a measure of location and estimates the population mean (expected value) $\mu = E(Y)$. The sample mean is often described as the "balance point" of the data. The following alternative description is also useful. For any value m consider the data values $Y_i \le m$, and the values $Y_i > m$. Suppose that there are n rods where rod i has length $|r_i(m)| = |Y_i - m|$ where $r_i(m)$ is the ith residual of m. Since $\sum_{i=1}^n (Y_i - \overline{Y}) = 0$, $\overline{Y}$ is the value of m such that the sum of the lengths of the rods corresponding to $Y_i \le m$ is equal to the sum of the lengths of the rods corresponding to $Y_i > m$. If the rods have the same diameter, then the weight of a rod is proportional to its length, and the weight of the rods corresponding to the $Y_i \le \overline{Y}$ is equal to the weight of the rods corresponding to $Y_i > \overline{Y}$. The sample mean is drawn towards an outlier since the absolute residual corresponding to a single outlier is large.

If the data $Y_1, ..., Y_n$ is arranged in ascending order from smallest to largest and written as $Y_{(1)} \le \cdots \le Y_{(n)}$, then $Y_{(i)}$ is the ith order statistic and the $Y_{(i)}$'s are called the *order statistics.* Using this notation, the median

$$\text{MED}_c(n) = Y_{((n+1)/2)} \text{ if n is odd,}$$

and

$$\text{MED}_c(n) = (1-c)Y_{(n/2)} + cY_{((n/2)+1)} \text{ if n is even}$$

for $c \in [0,1]$. Note that since a statistic is a function, c needs to be fixed. The *low median* corresponds to $c = 0$, and the *high median* corresponds to $c = 1$. The choice of $c = 0.5$ will yield the sample median. For example, if the data $Y_1 = 1, Y_2 = 4, Y_3 = 2, Y_4 = 5$, and $Y_5 = 3$, then $\overline{Y} = 3$, $Y_{(i)} = i$ for $i = 1, ..., 5$ and $\text{MED}_c(n) = 3$ where the sample size $n = 5$.

Definition 1.2. The *sample median*

$$\text{MED}(n) = Y_{((n+1)/2)} \text{ if n is odd,} \tag{1.3}$$

$$\text{MED}(n) = \frac{Y_{(n/2)} + Y_{((n/2)+1)}}{2} \text{ if n is even.}$$

The notation $\text{MED}(n) = \text{MED}(Y_1, ..., Y_n)$ will also be used.

Definition 1.3. The *sample variance*

$$S_n^2 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1} = \frac{\sum_{i=1}^n Y_i^2 - n(\overline{Y})^2}{n-1}, \tag{1.4}$$

and the *sample standard deviation* $S_n = \sqrt{S_n^2}$.

The sample median need not be unique and is a measure of location while the sample standard deviation is a measure of scale. In terms of the "rod analogy," the median is a value m such that at least half of the rods are to the left of m and at least half of the rods are to the right of m. Hence the number of rods to the left and right of m rather than the lengths of the rods determine the sample median. The sample standard deviation is vulnerable to outliers and is a measure of the average value of the rod lengths $|r_i(\overline{Y})|$. The sample mad, defined below, is a measure of the median value of the rod lengths $|r_i(\text{MED}(n))|$.

Definition 1.4. The *sample median absolute deviation* is

$$\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|,\ i = 1, \ldots, n). \tag{1.5}$$

Since $\text{MAD}(n)$ is the median of n distances, at least half of the observations are within a distance $\text{MAD}(n)$ of $\text{MED}(n)$ and at least half of the observations are a distance of $\text{MAD}(n)$ or more away from $\text{MED}(n)$.

Example 1.1. Let the data be $1, 2, 3, 4, 5, 6, 7, 8, 9$. Then $\text{MED}(n) = 5$ and $\text{MAD}(n) = 2 = \text{MED}\{0, 1, 1, 2, 2, 3, 3, 4, 4\}$.

Since these estimators are nonparametric estimators of the corresponding population quantities, they are useful for a very wide range of distributions.

1.5 Mixture Distributions

Mixture distributions are often used as outlier models, and certain mixtures of elliptically contoured distributions have an elliptically contoured distribution. The following two definitions and proposition are useful for finding the mean and variance of a mixture distribution. Parts a) and b) of Proposition 1.2 below show that the definition of expectation given in Definition 1.6 is the same as the usual definition for expectation if Y is a discrete or continuous random variable.

Definition 1.5. The distribution of a random variable Y is a *mixture distribution* if the cdf of Y has the form

$$F_Y(y) = \sum_{i=1}^{k} \alpha_i F_{W_i}(y) \tag{1.6}$$

where $0 < \alpha_i < 1$, $\sum_{i=1}^{k} \alpha_i = 1$, $k \geq 2$, and $F_{W_i}(y)$ is the cdf of a continuous or discrete random variable W_i, $i = 1, ..., k$.

Definition 1.6. Let Y be a random variable with cdf $F(y)$. Let h be a function such that the expected value $Eh(Y) = E[h(Y)]$ exists. Then

$$E[h(Y)] = \int_{-\infty}^{\infty} h(y) dF(y). \tag{1.7}$$

Proposition 1.2. a) If Y is a discrete random variable that has a pmf $f(y)$ with support $\mathcal{Y}$, then

$$Eh(Y) = \int_{-\infty}^{\infty} h(y) dF(y) = \sum_{y \in \mathcal{Y}} h(y) f(y).$$

b) If Y is a continuous random variable that has a pdf $f(y)$, then

$$Eh(Y) = \int_{-\infty}^{\infty} h(y) dF(y) = \int_{-\infty}^{\infty} h(y) f(y) dy.$$

c) If Y is a random variable that has a mixture distribution with cdf $F_Y(y) = \sum_{i=1}^k \alpha_i F_{W_i}(y)$, then

$$Eh(Y) = \int_{-\infty}^{\infty} h(y)dF(y) = \sum_{i=1}^{k} \alpha_i E_{W_i}[h(W_i)]$$

where $E_{W_i}[h(W_i)] = \int_{-\infty}^{\infty} h(y)dF_{W_i}(y)$.

Example 1.2. Proposition 1.2c implies that the pmf or pdf of W_i is used to compute $E_{W_i}[h(W_i)]$. As an example, suppose the cdf of Y is $F(y) = (1-\epsilon)\Phi(y) + \epsilon\Phi(y/k)$ where $0 < \epsilon < 1$ and $\Phi(y)$ is the cdf of $W_1 \sim N(0,1)$. Then $\Phi(y/k)$ is the cdf of $W_2 \sim N(0,k^2)$. To find EY, use $h(y) = y$. Then

$$EY = (1-\epsilon)EW_1 + \epsilon EW_2 = (1-\epsilon)0 + \epsilon 0 = 0.$$

To find EY^2, use $h(y) = y^2$. Then

$$EY^2 = (1-\epsilon)EW_1^2 + \epsilon EW_2^2 = (1-\epsilon)1 + \epsilon k^2 = 1 - \epsilon + \epsilon k^2.$$

Thus VAR$(Y) = E[Y^2] - (E[Y])^2 = 1 - \epsilon + \epsilon k^2$. If $\epsilon = 0.1$ and $k = 10$, then $EY = 0$, and VAR$(Y) = 10.9$.

To generate a random variable Y with the above mixture distribution, generate a uniform (0,1) random variable U which is independent of the W_i. If $U \leq 1-\epsilon$, then generate W_1 and take $Y = W_1$. If $U > 1-\epsilon$, then generate W_2 and take $Y = W_2$. Note that the cdf of Y is $F_Y(y) = (1-\epsilon)F_{W_1}(y) + \epsilon F_{W_2}(y)$.

Remark 1.1. Warning: Mixture distributions and linear combinations of random variables are very different quantities. As an example, let

$$W = (1-\epsilon)W_1 + \epsilon W_2$$

where W_1 and W_2 are independent random variables and $0 < \epsilon < 1$. Then the random variable W is a linear combination of W_1 and W_2, and W can be generated by generating two independent random variables W_1 and W_2. Then take $W = (1-\epsilon)W_1 + \epsilon W_2$.

If W_1 and W_2 are as in the previous example then the random variable W is a linear combination that has a normal distribution with mean

$$EW = (1-\epsilon)EW_1 + \epsilon EW_2 = 0$$

and variance

$$\mathrm{VAR}(W) = (1-\epsilon)^2 \mathrm{VAR}(W_1) + \epsilon^2 \mathrm{VAR}(W_2) = (1-\epsilon)^2 + \epsilon^2 k^2 < \mathrm{VAR}(Y)$$

where Y is given in the example above. Moreover, W has a unimodal normal distribution while Y does not follow a normal distribution. In fact, if $X_1 \sim N(0,1)$, $X_2 \sim N(10,1)$, and X_1 and X_2 are independent, then $(X_1+X_2)/2 \sim N(5, 0.5)$; however, if Y has a mixture distribution with cdf

$$F_Y(y) = 0.5 F_{X_1}(y) + 0.5 F_{X_2}(y) = 0.5\Phi(y) + 0.5\Phi(y-10),$$

then the pdf of Y is bimodal.

1.6 Summary

1) Given a small data set, find $\overline{Y}$, S, MED(n) and MAD(n). Recall that $\overline{Y} = \frac{\sum_{i=1}^n Y_i}{n}$ and the *sample variance*

$$\mathrm{VAR}(n) = S^2 = S_n^2 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1} = \frac{\sum_{i=1}^n Y_i^2 - n(\overline{Y})^2}{n-1},$$

and the *sample standard deviation* (SD) $S = S_n = \sqrt{S_n^2}$.

If the data $Y_1, ..., Y_n$ is arranged in ascending order from smallest to largest and written as $Y_{(1)} \leq \cdots \leq Y_{(n)}$, then the $Y_{(i)}$'s are called the *order statistics.* The *sample median*

$$\mathrm{MED}(n) = Y_{((n+1)/2)} \quad \text{if n is odd,}$$

$$\mathrm{MED}(n) = \frac{Y_{(n/2)} + Y_{((n/2)+1)}}{2} \quad \text{if n is even.}$$

The notation MED(n) = MED($Y_1, ..., Y_n$) will also be used. To find the sample median, sort the data from smallest to largest and find the middle value or values.

The *sample median absolute deviation*

$$\mathrm{MAD}(n) = \mathrm{MED}(|Y_i - \mathrm{MED}(n)|,\ i = 1, \ldots, n).$$

To find MAD(n), find $D_i = |Y_i - \mathrm{MED}(n)|$, then find the sample median of the D_i by ordering them from smallest to largest and finding the middle value or values.

1.7 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

1.1. Consider the data set 6, 3, 8, 5, and 2. Show work.

a) Find the sample mean $\overline{Y}$.

b) Find the standard deviation S

c) Find the sample median MED(n).

d) Find the sample median absolute deviation MAD(n).

1.2*. The Cushny and Peebles data set (see Staudte and Sheather 1990, p. 97) is listed below.

```
1.2  2.4  1.3  1.3  0.0  1.0  1.8  0.8  4.6  1.4
```

a) Find the sample mean $\overline{Y}$.

b) Find the sample standard deviation S.

c) Find the sample median MED(n).

d) Find the sample median absolute deviation MAD(n).

e) Plot the data. Are any observations unusually large or unusually small?