1 Introduction

1.1 Key Terms

This section will not be covered in class but you are expected to know it thoroughly. Please ask questions in class or come to office hours if you need help.

2 What is Multivariate Analysis?

This is the information age. What has become a truism in popular culture poses real difficulties for the researcher. Data have become so numerous and so complex that often even the best human minds have difficulty processing the results of even one experiment or study. When a large number of variables are measured it is tempting to try to perform a large number of pair-wise comparisons to make sense of the data, or ignore the experimental design and try to deal with the data as if they came from several rather than one experiment. However, we end up with so many comparisons and so many different analyses that we have merely delayed rather than avoided the task of understanding the complexity in our data. Multivariate analysis is a set of analytical tools for summarizing and simplifying data wherein several variables have been measured on one observational unit, so that we can detect patterns and significant relationships and thereby comprehend the results of our studies and experiments.

3 Basic Concepts of Multivariate Analysis

3.1 The Variate

A key concept in multivariate analysis is taking linear combinations of the data that simplify or highlight important patterns. These linear combinations are referred to by your text as variates. They are also often referred to as components or factors in the literature. A linear combination of \( p \) variables can be represented by

\[
V = w_1X_1 + w_2X_2 + \cdots + w_nX_p
\]  

The weights, \( w_i \), are chosen depending on the goal of the analysis. For example, in cluster analysis the weights are chosen (algorithmically) so as to maximally differentiate among the different groups; in principal components, the weights are chosen such that the variate ‘explains’ as much of the variation as possible (more later). In any case, the nature of the variate is both determined by and is the defining characteristic of the multivariate analysis being used.
3.2 Measurement scales

There are two main types of measurements that we typically encounter: metric and non-metric. However, classifying any given set of measurements as either metric or non-metric is not always straightforward. We will consider several types of variables below.

3.2.1 nominal or categorical variables

These are considered non-metric variables. For nominal or categorical variables the ‘measurement’ on the experimental unit consists of identifying the class to which the experimental unit belongs. There is no inherent scale to these observations, that is, we can’t say that one category is in any mathematical sense, less than or greater than the other. For example, ‘sex’ is a categorical variable. In most analyses we will assign a numeric value such as 0 = male, 1 = female, or 1 = dead, 0 = alive, etc. We could use any number to identify the classes. The use of 0 and 1 is conventional for binary data (data with two possible categories).

3.2.2 ordinal variables

Ordinal variables can be considered both a special kind of categorical variable and a special kind of metric variable. Here, the data are classified into categories but there is an inherent order to the categories. For example, the Mohs Hardness Scale classifies minerals and rocks according to 10 levels of hardness. A diamond could be rated at 10, whereas talc could be rated at 1, though any 10 numbers could be used in this scale so long as they are ordered according to magnitude. Hence there is an underlying continuum that is being represented by artificial categories. However, the scale of the continuum is not fully represented. For example, if diamond = 10, iron = 9, fluorite = 4 and calcite = 3, we cannot conclude that the difference in hardness between diamonds and iron is the same as the difference in hardness between fluorite and calcite, because there is a range of hardness that falls into any one category.

Ordinal variables are particularly appropriate when the precision of our measuring instrument is low. For example, if we are measuring percent ground cover of a particular species in meter plots by eye-balling, it may be more honest to classify the ground cover as ‘high’, ‘medium’, and ‘low’ because we really cannot reliably tell the difference between 20% and 17%. Hence recording percentages might be misleading, implying that a more precise analytical tool was used to obtain the measurement than actually was.
3.2.3 interval variables

An interval variable has the quality that unit differences anywhere on the scale are the same. So for example, on a temperature scale, the difference between 12° and 13° is the same as the difference between 214° and 215°. Note that this is not the case for ordinal variables.

3.2.4 ratio variables

Ratio variables are interval variables with a defined, natural zero point. They are called ratio variables because the value of ratios of these variables does not change with a change in units. This characteristic allows us to preserve ratio values across the scale. For example, a tree 20 feet tall is twice as tall as a tree measuring 10 feet. If we express the height in meters, then we have \( \frac{6.10}{3.05} = 2 \). However, we cannot say that a liquid that is at 200°F is twice as hot as a liquid that is at 100°F because if we transform the measurement from Fahrenheit to Celsius the value of the ratio does not remain equal to 2.

An interesting relationship between ratio and interval variables is that the difference between two interval variables is a ratio variable, because now there is a natural zero point, that is the point of no difference. For example, ‘time of day’ is an interval variable while ‘length of time’ (the difference between the start time and the end time) is a ratio variable.

3.2.5 continuous versus discrete

You will also hear the terms continuous and discrete to describe different sorts of variables. A continuous variable is one that can take on an uncountably infinite number of values. A discrete variable is one that can take on only a finite or countably infinite number of values. For example, weight can take on uncountably many values because (so long as our scale has infinite accuracy) there are an infinite number of weights between 130 lbs and 131 lbs, or any two numbers. However, the number of birds that visit a feeder in a 24 hour period is discrete because the counts can take on only integer values.

**Question**: Identify the type of variable (nominal, ordinal, interval or ratio).

1. IQ score
2. Amount of money in your checking account
3. Intensity of a color
4. Employment Status
5. Socioeconomic status
6. Species name
7. Depression score
4 Measurement Error and Multivariate Measurement

Please review the discussion of validity and reliability on your own. Feel free to ask questions in class or during office hours.

4.1 Summated Scales

In multivariate methods the information in the data is used by taking linear combinations (creating variates) in hopes that most of the useful information will be retained in fewer variates. That is, we reduce the dimension of the data without losing information by using a very few linear combinations of all the variables rather than all of the variables themselves. One major benefit of dimension reduction is that most classical statistical models require that the sample size be larger than the number of explanatory variables. Multivariate techniques allow us to include a large number of predictors even for small sample sizes. Often, the new variates will be scientifically meaningful and appealing. When this happens, we often consider the new variate to be an index of some quantity we are interested in. For example, scientists often create 3 or 4 ‘optical indices’ by calculating linear combinations of reflectance measurements at several hundred wavelengths of light. These indices are thought to indicate some overall condition of the plant with respect to photosynthetic and other activity, etc. Once this is achieved, different statistical analyses can be performed using the 3 or 4 variates rather than the several hundred.

Sometimes these linear combinations are developed using scientific reasoning, but they can also be developed using multivariate techniques that optimize some mathematical criterion. Often the linear combination found using the algorithm is just as mechanistically appealing as that found using scientific reasoning (though not always).

5 Statistical Significance versus Statistical Power

This section will not be covered in class but you are expected to know it thoroughly. Please come to office hours if you have questions.

6 Types of Multivariate Techniques: a quick survey

Deciding which multivariate technique to use can be challenging. Below we will consider several examples of multivariate data as well as brief descriptions of the possible approaches.

Variables in an analysis can often be referred to as ‘response’ or ‘dependent’ versus ‘predictor’, ‘independent’ or ‘explanatory’. I will tend to use the terms interchangeably and use ‘response’ and ‘predictor’ most often. Your text uses ‘dependent’ and ‘independent’ almost exclusively. Not all data sets deal with prediction dependence, and so the variables cannot be classified in this way. In some cases, we are exclusively interested in interdependence. Be thinking of this issue as we go through examples. Which data sets are concerned with
prediction, which with interdependence? If prediction is the goal, which variables are the response and which are the explanatory?

6.1 Examples of Multivariate Data

Example 1 Bumpus’s Sparrow Data
Hermon Bumpus, a professor at Brown University in Rhode Island, collected several moribund sparrows after a severe storm on 1 February 1898. Bumpus saw this as an opportunity to test Darwin’s Theory of Natural Selection by looking to see if the birds that subsequently died were physically different from those that lived. He made 8 morphological measurements on each bird, such as lengths of different bones, and also weighed them. From the data he concluded that those birds that died did so because they were physically less fit than those who survived. Bumpus did not have access to even our most simple statistical techniques but later investigators have confirmed his conclusions.

To illustrate multivariate techniques, let’s consider the possible questions one could ask of the data:

1. How are the various measurements related?
2. Do the survivors and non-survivors have statistically different mean values for the variables? If so, which ones?
3. Do the survivors and nonsurvivors show similar amounts of variation for the variables?
4. If there are differences, is it possible to construct some function of the variables that separates the two groups?
5. Can we find an appropriate ‘index’ of Darwinian fitness as a linear combination of the variables?

Example 2 Distribution of a Butterfly
A study of 16 colonies of the butterfly Euphydrus editha in California and Oregon was conducted by McKechnie et al in 1975. Four environmental variables were measured: altitude, annual precipitation, max and min temperatures, and six genetic variables: percent frequencies for different Pgi genes as determined by electrophoresis.

Questions that could be asked:

1. Are the Pgi frequencies similar for the colonies that are close in space?
2. To what extent, if any, are the Pgi frequencies related to the environmental variables?

Example 3 Employment in European Countries
The data on percent employed in nine different industries for 30 different European countries were collected by the U.N. Department of Social Affairs in 2000.

Questions:
1. Can the countries be classified by employment patterns across the 9 industries?
2. What are the relationships between the countries with respect to employment patterns?
3. Are the differences between countries related to political grouping, (EU, EFTA, eastern europe, other)?

6.2 Types of Multivariate Techniques

Now let’s consider the different types of MV techniques and when and where they might be useful.

6.2.1 Principal Components, Common Factor, and Partial Least Squares Analysis

These are considered data dimension reduction techniques that produce summed scales. These techniques are useful in elucidating relationships between variables. Typically, the variables are all metric.

Example 4 PCA: Forced Vital Capacity

An investigator has made a number of measurements of lung function on a sample of adult males who do not smoke. These measurements consist of forced vital capacity, forced mid-expiratory flow, maximal expiratory flow at 50% forced vital capacity, and other measures of lung function, all measured using the trace of the volume of air expired over time through a spirometer. Since all these measurements are made from the same volume of air, they are highly interrelated. From past experience it is known that some are more related than others and that they measure airway resistance in different sections of the airway.

The investigator can perform a principal components analysis to determine whether the variates derived from the PCA can explain most of the variation in the original lung function measurements. In PCA, the resulting components are all uncorrelated and explain as much of the variation as possible. It is also hoped that some operational meaning can be ascertained from the components that will aid in interpretation. Also, the components can now be used in a linear regression to predict performance in cardiovascular exercise.

Example 5 FA: Attitudes towards Inflation

An investigator performs a survey where each respondent is asked whether he or she strongly agrees, agrees, is undecided, disagrees, or strongly disagrees with 15 statements about inflation. As a first step, the investigator can perform a factor analysis to determine which statements belong together in uncorrelated sets. The particular statements that are important in forming the sets will be examined to obtain a better understanding of attitudes towards inflation. Scores derived from each set or factor can be used in subsequent analyses to examine how attitudes towards inflation predict voting behavior.

Example 6 PLS: Hyperspectral reflectance and contamination

An investigator measures the reflectance at 1000 wavelengths on a species of plants growing in
contaminated and uncontaminated areas within a salt marsh near a petroleum refinery. She would like to classify the plants as contaminated or uncontaminated based on their reflectance spectra. She performs a partial least squares analysis to reduce the dimension of the data to 4 or 5 linear combinations of the wavelengths (components). She then performs a logistic regression using the components to predict contaminated or uncontaminated.

6.2.2 Canonical Correlation

Canonical correlation involves both multiple response and multiple predictor variables. The goal of this analysis is to find the linear combination of response variables (canonical response variable) that is maximally correlated with a linear combination of the predictor variables (canonical predictor variable).

**Example 7** Depression and Physical Well-being

A psychiatrist has data consisting of depression and physical well-being scores and ‘social’ data on age, sex, income, number of social contacts per month, and marital status. Using canonical correlation, the psychiatrist will determine the canonical variable of the social data that is most highly correlated with the canonical variable of the depression and physical well-being scores.

6.2.3 Multiple Discriminant Analysis

This technique is used to classify subjects into two or more groups using several independent variables. The main objective is to understand group differences and estimate the probability of the subject belonging to a particular group based on the independent variables. Here the response is categorical and the predictors are usually metric, though categorical predictors can also be used.

**Example 8** Heart Disease

A large sample of initially disease-free men over 50 years of age has been followed to see who subsequently develops heart disease. At the first visit, several blood tests are performed such as serum cholesterol, phospholipids, blood glucose, as well as body mass index and others. Ten years later, the investigator observes the men who did and did not have heart attacks. The investigator would like to determine a linear function of these variables that would be useful in predicting who will experience a heart attack within 10 years.

6.2.4 Multivariate Analysis of Variance

MANOVA is a direct extension of univariate ANOVA. Here we measure several continuous variables on each subject and test simultaneously whether there is a treatment effect on this multivariate response. If the experimental design involves multiple response variables, it is inappropriate to simply perform multiple univariate ANOVAs. Hence MANOVA is important in these studies. In MANOVA the response is metric and the predictors are categorical.
Example 9 Effects of Human Disturbance of Reproduction

An investigator is interested in the effects of slash and burn agricultural practices on subsequent recovery of the rain forest in New Guinea. He measures several reproductive variables on several individuals of an endangered plant species that grows in abandoned agricultural fields. He measures the number of seeds produced, the weight of seeds produced, number of seeds germinating, and number of seedlings surviving after one year in several randomly selected abandoned agricultural fields that can be classified into two types. In the first type, minimal wild vegetation was cleared by hand and the crop planted around the existing large trees. The second type was cleared by first cutting the vegetation and leaving it to dry. After several days, the dried vegetation was burned. Both fields were farmed for 3 years before coming under protection by an environmentally responsible oil company who purchased the land. Both types of fields are found on steep slopes and so he blocks the fields into three elevation ranges. A MANOVA will test simultaneously whether any of the response variables differ for the two farming practices across the three elevations.

6.2.5 Cluster Analysis

Cluster analysis is used to discover previously unknown clusters. It differs from discriminant analysis in that the number of clusters as well as cluster membership is unknown before the analysis is performed. Here the variables are usually metric.

Example 10 Depression

Investigators have made numerous measurements on a sample of patients who have been classified as being depressed. They wish to determine whether the patients can be classified into natural groupings that can be identified as distinct types of depression based on the patients scores on various diagnostic tests.

7 A Classification of Multivariate Techniques

Please study this section and refer to the diagram on pages 14 and 15. The diagram below shows a slightly different classification of MV techniques. Note that the exploratory methods can be used to create variates that can themselves be used in inferential techniques, such as regression.

These figures may be helpful to you in your future researches as you both design experiments and analyze the resulting data or data from observational studies. Keep them in the back of your mind to help you organize what you learn throughout the semester.
8 Guidelines for Multivariate Analyses and Interpretation

Because multivariate techniques are diverse and powerful, it can be tempting to believe that we can skip the hard scientific work in an experiment. Why bother coming up with elegant mechanistic theories to test, when we can simply measure every variable we can think of and throw the data into a multivariate analysis and see what pops out? The answer is that, ultimately, this approach to science is no fun and usually not very productive. The above approach is appropriate only when we are encountering a virgin system that no one knows anything about. This almost never happens. If it happens to you, you should consider yourself extraordinarily lucky. You should always do a thorough literature review and spend some time thinking hard about what it is you really want to know and developing a scientific model that you hope to confront with the data.

8.1 Establish Practical Significance

With large samples or with a large number of variables, there’s a potential that even very small effect sizes can be detected. If an effect is very small, is it still interesting? This is a question of practical significance. How small an effect is interesting is a question that should be answered before the experiment is designed or the data collected. This question can only be answered by the scientist.

8.2 Sample Size Affects All Results

It is best to estimate what size sample is required before the data are collected. However, sample size calculations and a priori power analyses can be quite difficult for complicated statistical procedures because the distribution of the test statistic can be ugly or unknown. Unequal sample sizes can also affect multivariate procedures because they involve optimization. An example of this occurred in my own research using a multivariate technique called ‘Support Vector Machine Classification’. The technique seeks to minimize prediction error. In one of my data sets, the uncontaminated plants were far fewer than the contaminated plants. The data were quite noisy and the procedure minimized the prediction error by simply classifying everything contaminated (only 25% error!). All results should be interpreted in light of the sample size.

8.3 Know Your Data

We will cover a variety of exploratory and diagnostic techniques for multivariate data. Always use them!

8.4 Strive for Model Parsimony

The inclusion of too many variables results in the phenomenon of ‘over-fitting’. If we include enough variables we could predict perfectly all the observations in our data set. However, if
we were to then try to predict new data, not included in the model-fitting procedure, most likely, our prediction error would be horrible. On the other hand, not including important variables lead to spurious correlations and bad decision-making. For these reasons, as well as others, measuring everything we can think of and throwing it all into a multivariate procedure is a very bad idea. If we are spending our time thinking only of what can be measured and not of what should be measured, then we are failing at our jobs as scientists.

8.5 Look at Your Errors
The troublesome data points are often the most interesting from a scientific point-of-view. Residual analysis in regression settings is important for model validation and assessment. Looking at the misclassified points in a discriminant analysis or the outliers in a cluster analysis can help us identify how to improve the model and the issues that we may heretofore been neglecting.

8.6 Validate Your Results
We will cover a couple of cross validation techniques in this course. Always use them! Cross validation is the best way to avoid over-fitting. Cross validation techniques give us an estimate of the predictive power of the model. Even if you are not interested in prediction, but only in elucidating underlying mechanisms, the best way to evaluate the validity of the mechanistic model is to evaluate how predictive it is.

9 A Structured Approach to Multivariate Model Building

- **Stage 1**: Define the research problem, objectives, and multivariate techniques to be used.
- **Stage 2**: Develop the analysis plan.
- **Stage 3**: Evaluate the assumptions underlying the multivariate technique.
- **Stage 4**: Estimate the multivariate model and assess model fit.
- **Stage 5**: Interpret the variates.
- **Stage 6**: Validate the model.

10 References