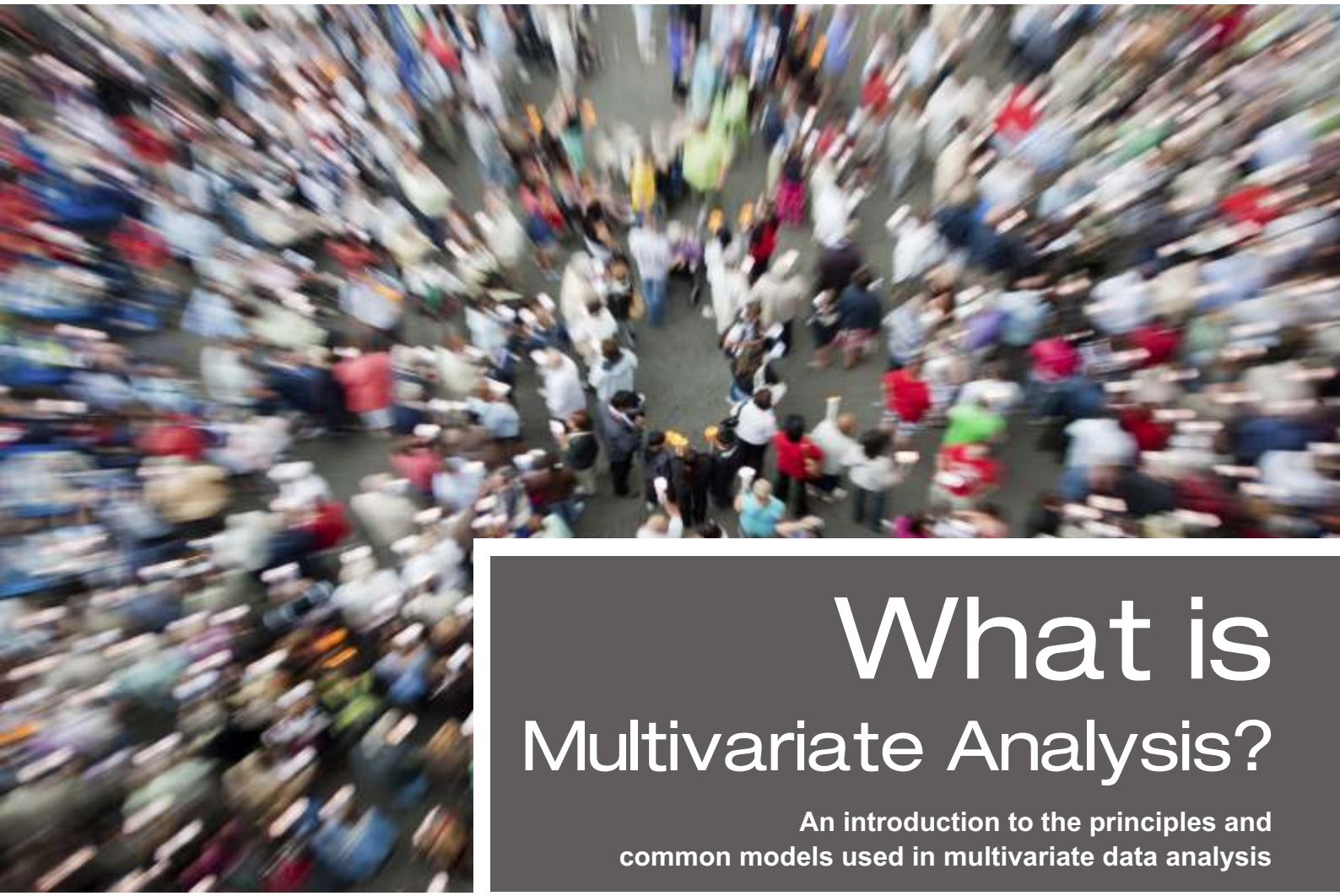
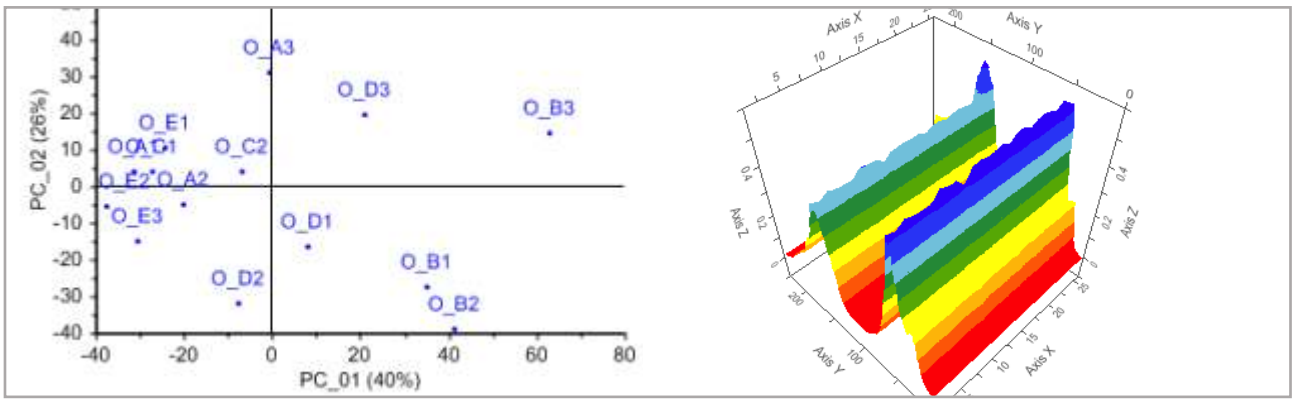


What is Multivariate Analysis?

An introduction to the principles and common models used in multivariate data analysis

camo.com

Bring data to life



Introduction

Most of the problems in the world are multivariate in nature - meaning that there are many variables that contribute to them. We cannot simply use the season of the year to predict the weather, as many other variables are part of the relationship to weather. The same holds true with other scientific, economic, and consumer preference relationships, such as product quality, which is influenced by numerous process variables, or consumer preference for a product which can be based on traits such as color, taste and price.

This short guide gives you an introduction to the principles of multivariate analysis, some broad applications for this technology, how it differs from classical (univariate) statistics and an overview of common multivariate models.

What is Multivariate Analysis?

Imagine out of the five senses you only had sight. From your perspective you could see the world but you would not be able to hear the sounds around you, smell, taste or feel things. Your understanding of the world would be more limited.

In the event of danger, provided you were pointed in the direction of the danger, you would have some chance of avoiding it. But if it was behind you, you would never hear the danger, or know that it was coming. With the combination of sight and sound it is easier to avoid danger. Even with two of the five senses, your view of the world is still limited.

Multivariate analysis is a tool to find **patterns** and **relationships** between several variables simultaneously.

Most of us use all of our senses to understand the world around us i.e. not just one “measurement” but the combination of several senses working together. This is like multivariate analysis. In multivariate analysis we use the information from many sources simultaneously to get a better picture of our surroundings.

Essentially, multivariate analysis is a tool to find patterns and relationships between several variables simultaneously. It lets us predict the effect a change in one variable will have on other variables.

Again using the example of our senses, while taste and smell are two separate measures, they are not independent of each other. For example, if it smells bad, it often tastes bad. This gives multivariate analysis a decisive advantage over other forms of analysis.

Multivariate analysis is also highly graphical in its approach. This allows an analyst to examine the inner or hidden structure of large data sets, and to visually identify the factors which influence the results. The expression ‘a picture is worth a thousand words’ is particularly relevant when trying to interpret large, complex data sets!

Multivariate analysis is used widely in many industries, from raw material analysis and drug discovery in the pharmaceutical industry, early event detection and gasoline blending in refineries, right through to predicting future market trends in business intelligence applications. It can be used for measuring data sets with many input variables or for investigating the trends in time series data, all of which provide a better understanding of a given issue and often result in resource and time savings.

Using one more example, if we consider the rate of fuel used by a car (based on the time it takes for the indicator to go from full to empty), this is not just a simple function of driving the car. It is a function of the size of the engine, whether the car was driven on flat or mountainous roads, the load in the car, whether the car is in a good state of repair, the grade of gasoline used etc. To fully understand this problem, a number of cars would have to be tested under varying conditions and analysed using multivariate methods to reveal the patterns and relationships. From there a model could be developed to help select the right car for the right conditions. Wherever many measurements on many variables are performed, this data is ideal for multivariate analysis.

For most real-world problems we need to use multivariate analysis to model the relationship to the response. We need tools to mine our data, and understand the relationships in them, whether for pattern recognition, or to develop models that can be used to predict values for new samples.

Multivariate analysis (MVA) has wide application to data including instrumental data, medical diagnostics, census data, economic data, marketing data, or even a sports team's performance. MVA gives us a means to find the relationships in the data, and provides tools to visualize the relationships between samples and variables. It can be used for both qualitative and quantitative analysis.

How are multivariate methods used?

MVA methods are used for:

- Pattern recognition by providing an understanding of the important patterns & underlying relationships in data
- Getting deeper insights into data allowing you to model and visualize complex data for more insightful analysis
- Predicting behavior and improving forecasting of likely outcomes using predictive tools

Important decisions and products we use every day are often based on **Univariate analysis** that does not necessarily tell the full story or might not even be accurate.

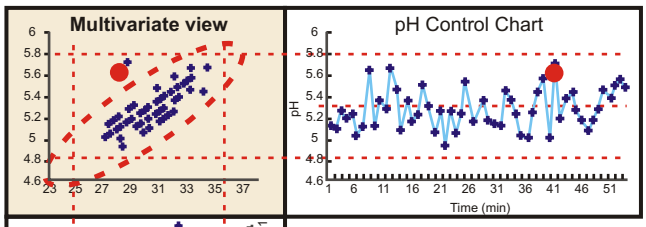
Multivariate analysis versus univariate analysis (classic statistics)

Most people have heard of the mean, median, standard deviation, normal distribution etc. These are univariate - or classical - statistics. Univariate statistics can be useful, but are limited by only looking at one variable at a time. In reality, there's often more than one variable involved, so univariate analysis can lead to the wrong conclusions. It is often necessary to sample, observe, study or measure more than one variable simultaneously to understand a process or any set of samples with numerous measurements.

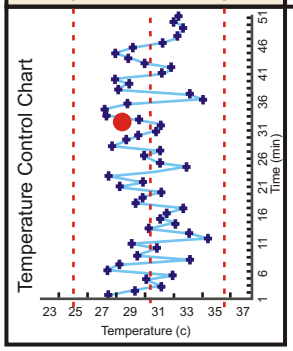
An illustration of why this is important is shown in the figure below, where the univariate plots of two variables, pH and temperature are shown along with a multivariate plot of these two variables. Because there is a correlation between these univariate measurements, the univariate plots do not capture the fact that there is a sample for which there is a problem in the relationship between the pH and temperature values, shown by the red dot.

Multivariate analysis provides a more accurate depiction of the behavior of data that are highly correlated, and can indicate when there are potential problems in a system or process. While this might not seem to be a big problem, in many situations important decisions and products we use every day are often based on univariate analysis that does not necessarily tell the full story or might not even be accurate.

This example also illustrates how important visualization of data is. Making various plots of our data helps us to see trends and relationships that are not evident when looking at lists or tables of numbers.



Univariate view of variable 1



Univariate view of variable 2

- *Only with multivariate analysis can the fault be detected*
- *The univariate limits are too wide to detect a multivariate fault*
- *The two variables under consideration are not independent*
- *The "sweet spot" is defined by the ellipse*
- *You can not see the outlier without looking at both variables at the same time*

Overview of common multivariate models

There are numerous tools used in multivariate analysis, from descriptive statistics to exploratory data analysis, and onwards to quantitative regression models. We often use descriptive statistics, principal component analysis, and clustering in our initial explorations of our data.

To make quantitative models, regression methods such as partial least-squares regression (PLSR) are used. Regression is used to develop a model using known samples and responses that can provide a predictive model. To identify or classify samples into groups with similar characteristics, there are various classification models such as Soft Independent Modeling by Class Analogy (SIMCA) or discriminant analysis.

One thing to keep in mind is that modeling is an iterative process. As we work though our data, we may need to remove outliers, focus on just some of the measurements, and make other adjustments to optimize our analysis and give us a picture that is descriptive of our system.

Depending on the objective of data analysis, multivariate data can be used to understand and model numerous outcomes. A summary of the different model types is given in the following table.

Descriptive model(s)	Regression & Predictive model(s)	Classification model(s)
Principal Component Analysis (PCA)	Multiple Linear Regression (MLR)	SIMCA (PCA, PLSR)
Basic Statistics	Principal Component Regression (PCR)	Support Vector Machine (SVM)
Clustering	Partial Least Squares Regression (PLSR)	Linear Discriminant Analysis (LDA)
		Partial Least Squares - Discriminant Analysis (PLS-DA)

To find out how multivariate analysis can be used in your industry, please visit www.camo.com or contact us for more information.

About CAMO Software

Founded in 1984, CAMO Software is a recognized leader in multivariate data analysis and Design of Experiments software and solutions. Our flagship software, The Unscrambler® X, is known for its ease of use, outstanding visualization and powerful analytical tools. More than 25,000 people in 3,000 organizations worldwide use our solutions for analysis and predictive modeling of complex data, saving time and money, and making better decisions based on more accurate information. Our in-house experts advise on data analysis across all industries as well as Process Analytical Technology (PAT) and Quality by Design (QbD) initiatives. For more information please visit **www.camo.com**

The Unscrambler® X
TRY ONLINE

CAMO

NORWAY
Nedre Vollgate 8,
N-0158
Oslo
Tel: (+47) 223 963 00
Fax: (+47) 223 963 22

USA
One Woodbridge Center
Suite 319, Woodbridge
NJ 07095
Tel: (+1) 732 726 9200
Fax: (+1) 973 556 1229

INDIA
14 & 15, Krishna Reddy
Colony, Domlur Layout
Bangalore - 560 071
Tel: (+91) 80 4125 4242
Fax: (+91) 80 4125 4181

AUSTRALIA
PO Box 97
St Peters
NSW, 2044
Tel: (+61) 4 0888 2007

JAPAN
Shibuya 3-chome Square Bldg 2F
3-5-16 Shibuya Shibuya-ku
Tokyo, 150-0002
Tel: (+81) 3 6868 7669
Fax: (+81) 3 6730 9539