# A Short Course in Multivariate Statistical Methods with R/Олон хэмжээст статистикийн богино хэмжээний сургалт R программ дээр

author: Ryan Womack, Rutgers University, [rwomack@rutgers.edu](mailto:rwomack@rutgers.edu) date: 2018-02-26 autosize: true

Based on Brian Everitt and Torsten Hothorn, *An Introduction to Applied Multivariate Statistical Analysis with R*, Springer, 2011.

First presented at Mongolian University of Life Sciences, February 2018

# Outline

* R environment, setup, basics
* Multivariate Analysis - what is it?
* Exploration and Visualization
* Principal Components
* Multidimensional Scaling
* Exploratory Factor Analysis
* Confirmatory Factor Analysis
  + Structural Equation Modeling
* Cluster Analysis
* Repeated Measures
* Additional topics, wrapup

# Goals

* Exposure to R
* Familiarity with major concepts used in multivariate analysis
* Implement these tools in R
* Learn “how to learn” - investigate and solve your own data problems
* Mastery is not possible in a short course. Don’t worry!

# the R environment

* **R**
  + free - easy to use and expand
  + open - fast and innovative
  + first - cutting edge of Data Science
* **R Markdown**
  + supports integration of code and text
  + multiple outputs (doc, html, pdf)
* **RStudio**
  + consistent, coding friendly developing environment
  + tools for publishing (Rpres, Rpubs)
  + literate programming
  + cross-platform and server version

For more details on authoring R presentations please visit <https://support.rstudio.com/hc/en-us/articles/200486468>.

# Setup

* download R
  + [R-project.org](https://r-project.org)
* download RStudio
  + [Rstudio.com](https://rstudio.com)
* code and presentation available from
  + [github.com/ryandata](https://github.com/ryandata/multivariate)
* [Rstudio.cloud](https://rstudio.cloud) is an experimental cloud server for R, free for now
* other texts distributed locally
* [YouTube.com/librarianwomack](https://youtube.com/librarianwomack) has tutorials. Other projects, papers, materials listed on [ryanwomack.com](https://ryanwomack.com) website

# Multivariate Data - what is it?

* for each subject, we have multiple variables and/or multiple observations
* if each variable is studied alone, the full structure of the data may not be revealed
* “multivariate statistical analysis is the simultaneous statistical analysis of a collection of variables, which improves upon univariate analysis…”
* Graphical methods and formal analysis will help us understand this data
* Computing has made multivariate methods more routine and widespread
* Cases with missing data can be excluded, or values can be imputed. See “hypo” data. This is a topic unto itself, so it is not treated further here.

# Covariance, Correlation, Distance, and the Multivariate Normal Distribution

* Basic statistical methods such as **Covariance** and **Correlation** are starting points for working with multivariate data
* **Distance**, usually Euclidean distance, is also commonly used
* The **Multivariate Normal Distribution** is most commonly assumed as a distribution of the underlying data, when this is required to advance the analysis.
  + The multivariate normal is “well-behaved” in roughly the same way that the univariate normal distribution is.

# Probability Plots

* We may need to test whether data fits the multivariate normal distribution
* If (MV) normal, distance metric of a single variable will have a chi-squared distribution
* Plot (computation illustrated by R code) should show data points roughly on a straight line
* Can identify outliers

# Data Visualization

* Advantages of visualization:
  + Easily detect patterns in data
  + Generate greater interest, understanding, and recall [for non-specialists and specialists alike]
  + Compress the meaning of large amounts of data into a smaller set of images
  + Discover hidden structure of data

# Basic Methods applied to Multivariate Data

* Scatterplot
* Bivariate Boxplot
  + alternatively, Convex Hull
* Chi-plot - should fluctuate around zero if independent

# Bubble and Glyph plots

* Bubble plot - size and shading of bubble introduces additional dimensions to the data
* Glyph plots - multidimensional, can be hard to interpret

# Scatterplot Matrix and Kernel Density

* Scatterplot matrix plots multiple variables against each other simultaneously
* Kernel density visualizes the distribution of data
* These two methods can be combined to create a powerful summary of multivariate data

# Three-dimensional data

* Many tools can be used to visualize data in three dimensions
* Just a few examples in the code, more are illustrated at [my Data Visualization workshop](https://github.com/ryandata/DataViz)

# Principal Components Analysis

With multivariate data, we have **too many variables** \* Exploratory data analysis by methods such as scatterplots quickly becomes difficult \* We need to reduce the number of variables under consideration \* Example: GPA (Grade point average) is used instead of a long list of individual grades in courses to summarize a students’ achievement \* This is just a (weighted) combination of variables

# PCA, continued

* The *principal components* in principal components analysis are vectors
* Each vector is a linear combination of variables
* We want to find the smallest number of vectors that account for most of the variation in the data
* We do not know beforehand which variables are most useful for this task
* PCA solves this problem
* A **low dimension summary** of the data for graphing or other representations

# Solving the problem

* In one dimension, this is the same as determining the line that best fits the data
* In *m* dimensions, we find the m-dimensional projection that best fits the data
* This is the projection determined by variables with non-zero eigenvalues

# Scale

* This method is not *scale-invariant*, i.e., it produces different results for different units of measurement
* So, studying the covariance matrix for solutions also faces the scale-invariance problem
* In practice, we use the correlation matrix instead to generate solutions (which is scaled to unit interval)
* This also means we are essentially assuming that all variables will be equally weighted, with equal potential of being part of the solution (not always appropriate)

# How many components?

* The components are directly related to the covariance matrix
* We can select the number of components that allows us to approximate S efficiently:
  + by setting a target coverage of S (80% of variance)
  + by setting a cutoff value for $latex $ (e.g. 0.7, or simply more than the average for the data)
  + by using a scree diagram (looking for a bend or “elbow”)

# Principal Components Scores

* The principal components score for each observation is not predictive of the outcome (like the predicted values of a regression)
* But it shows which components are influential for each observation
* Scaling the data is often recommended to make interpretation clearer and more reliable
* Extreme caution should be used when “labeling” resulting components with meaning. The mathematical explanation of variance does not imply causal relationships.

# Multidimensional Scaling

* An extension of PCA’s methodology
* Extract a low-dimensional representation of the data, preserving relative distances
* Works on the distance matrix
* Some measurement of how similar or dissimilar items are
* Here, two spatial methods:
  + *Classical Multidimensional Scaling*
  + *Non-metric Multidimensional Scaling*

# Solving MDS

* Start with the (Euclidean) distance matrix (sometimes all we have)
* Compute an estimate of original data
* Because this method also uses the eigenvalues that account for most of the variation, it is equivalent to principal components, and often called *principal coordinates*
* Find where $latex $ are “high”
* Minimum spanning tree (“mst” command from “ape” package) can identify close groupings of observations

# Non-metric MDS

* Typically with ordered or ranked data, we can use a non-metric technique
* *isoMDS* command from “MASS” package
* use Shepard diagram to diagnose fit

# Correspondance Analysis

* essentially a method for plotting associations between categorical variables
* Row variables that appear close in a plot are similar
* Column variables that appear close are similar
* Row/column pairs indicate association