# A Short Course in Multivariate Statistical Methods with R

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

# Exploratory Factor Analysis

* Latent variables cannot be measured directly
* Manifest variables are linked to the underlying latent variables
* E.g., intelligence may manifest in academic performance, test scores
* Factor analysis enables us to discover relationships between manifest and latent variables
* **Exploratory Factor Analysis** focuses on the discovery process
* Does not prove the strength of the relationship or test the model (Confirmatory Factor Analysis)

# The k-Factor Analysis Model

* *q* observed variables (the manifest variables)
* *m*<*q* latent variables, modeled by
* The a,b,c… coefficients are essentially regression coefficients, but are called **factor loadings** in factor analysis. The *u* is random disturbance, assumed to be uncorrelated.
* Therefore, correlation among observed variables arises only due to relationship with the common underlying latent variables. Not due to correlated errors.

# The k-Factor Analysis Model, continued

* Because latent variables, termed **factors**, are unobserved, we can arbitratily fix their scales and locations.
* We assume they are standardized with mean 0, standard deviation of 1
* Also assume factors are uncorrelated with each other => factor loadings are the correlations of manifest variables with factors
* Variance of observed variables can be split into *communality* and *specific* parts
* The communality for a manifest variable is the most useful to estimate factors

# Relationship to PCA

* We solve the model by using the observed data’s covariance matrix
* Uses similar technique as Principal Components
* But because scale is arbitrary, covariance and correlation matrix produce equivalent results
* Solution methods are either *principal factor analysis* or *maximum likelihood estimatation [MLE]* (details in text)
* Solution by iteration occasionally results in a *Heywood case* (negative variance)

# Finalizing the model

* Selection of the number of factors is easiest when using MLE
* standard $latex ^2 $ test of whether variance reduction is significantly improved by adding additional factor
* Also, solutions are not unique due to the arbitrariness of the underlying variables
* Typically use **factor rotation** to produce factor loadings that are either large and positive or zero. Called a “simple structure”
* **Caution!** Although factor analysis point distribution is invariant to rotation, this is NOT true of principal components analysis
* Finally, we can assign **factor scores** to each observation

# Explanatory Factor Analysis, conclusions

* EFA is a tool for understanding features of data
* *factanal* command is used in **R** to implement
* EFA is starting point for more rigorous investigation, modeling
* Overinterpretation of results and meaning of the latent variables has been heavily criticized
* The latent variables are, after all, hypothetical. Not the same as factually observed data

# Confirmatory Factor Analysis

* Using EFA, we arrive at a proposed model
* Is it valid?
* We must test it on **new** data
* **Confirmatory Factor Analysis** serves this purpose
  + A subset of the more general methodology, **Structural Equation Modeling**

# Modeling and Testing

* Again, we will use *Maximum Likelihood Estimation (MLE)* to test model fit
* We assume the multivariate normal distribution describes the data
* If the number of free parameters and proposed relationship equations are indeterminate, the model is **unidentifiable**
* One requirement, , where t is the number of free parameters to q manifest variables
* Otherwise, no simple rules for detemining identifiability

# Assessing Fit

* $latex ^2 $ statistic is typically used on the fitted vs. unconstrained covariance matrix
* Other measures, Goodness of Fit, Adjusted Goodness of Fit, and more
* normed residuals < 2 is another check

# Performing the Confirmatory Factor Analysis

* outline proposed equations
* set certain parameters to zero where we believe there is no relationship
* estimate remaining free parameters
* then test whether fit is good
* in R, *sem* package is used for this, and other kinds of structural equation modeling
* uses special model notation
* path diagram to explain final model

# Structural Equation Modeling

* CFA can be considered a kind of constrained Structural Equation Modeling (SEM)
* SEM allows any kind of relationship between manifest and latent variables to be proposed
* Can lead to complex and difficult to interpret models
* Best to base relationships on well-estabilished disciplinary knowledge
* SEM is a large and growing research area on its own

# Cluster Analysis

* Classification is a fundamental tool for understanding data, with application across physical, life, and social sciences
* Cluster analysis provides numerical methods for sorting data into meaningful groups.
* Many methods are possible, 3 are described here:
  + agglomerative hierarchical techniques
  + k-means clustering
  + model-based clustering

# Agglomerative hierarchical techniques

* Hierarchy is generated by steps
* Start with each individual observation
* Merge the closest two observations into a cluster
* Repeat…
* Relies on distance metric (often Euclidean)
* Methods vary (using max distance or min distance between clusters, or a central measure)

# Finding the optimal cut point

* *hclust* is the R function that implements hierarchical clustering
* Typically, we plot a *dendrogram* to represent the clustering
* We can cut the dendrogram at the point that represents the maximum change in height, which is equivalent to the most dramatic reduction in average distance between clusters
* No precise method for this, although principal components analysis can help us validate the choice of groups

# k-means clustering

* In general, minimize some metric of aggregate distance
* In practice, minimize the within group sum of squares by choosing optimal k

$$latex \Huge {WGSS=\sum\_{j=1}^{q}\sum\_{l=1}^{k}\sum\_{i \in G}(x\_{ij}-\bar{x}\_j^{(l)})^2} $$

* Iterative process that finds a local, but not necessarily global, minimum by moving one element at a time among clusters to see if it reduces group sum of squares

# k-means continued

* k-means imposes spherical clusters due to its method, even if a better-fitting, odd shaped cluster is available
* k-means is **not** scale invariant
* One way of finding optimal number of k is to plot the WGSS, and look for an “elbow”, or prominent angle in the plot. This indicates that WGSS no longer reduces signficantly from adding additional clusters.

# Model-based clustering

* A model for the population from which data is sampled provides more tools for selecting clusters via statistical criteria
* With subpopulations in different amounts and distributions, a *finite mixture density* for the population as a whole is generated
* Probabilities associated with subpopulations are estimated via Maximum Likelihood Estimation (usually via iterative method)
* *mclust* package in R implements this
* We can plot clusters in various ways, such as the “stripes” plot in package *flexclust*
* Stripes plot reveals overlap and separation between clusters across multiple dimensions

# Repeated measures

* Repeated measures describe some of the most common situations in data analysis
* Collecting multiple samples/observations on a single subject
* Collecting data over time
* Data can be recorded in “wide” or “long” format (convert with *reshape* command)

# Mixed-effects models

* We know that the repeated observations are related to each other
* So treating each observation as independent (e.g., standard linear regression) is not appropriate
* Model must separate the variation into two components’
* within group (of repeated measures)
* across groups

# Random Intercept Model

$$latex \Huge {y\_{ij} = (\beta\_0 + u\_i) + \beta\_1x\_j+\epsilon\_{ij} } $$

* Each subject has a different intercept ($latex u\_i $)
* Slope is common across all subjects
* Intercept is the “random effect”, composed of some common $latex $ combined with a subject-specific effect
* Slope is the “fixed effect”

# Random Intercept and Slope Model

$$latex \Huge {y\_{ij} = (\beta\_0 + u\_{i1}) + (\beta\_1 + u\_{i2})x\_j+\epsilon\_{ij} } $$

* Each subject has a different intercept ($latex u\_i latex /beta\_i $)
* Slope varies according to subject
* Both intercept and slope are random effects
* Can account for more complexity, variation

# Solving the model

* We use *restricted maximum likelihood estimation* to solve the model (regular MLE underestimates variances)
* *lme* command in **R** along with model specification
* Exact Likelihood Ratio Test (*exactLRT* from package *RLRsim*) to find p-value to test model against independent variable model
* LRT from ANOVA to test competing mixed effects models
* Model equations can be modified for improved fit (e.g. with quadratic terms), just like regular regression

# Wrapping Up

* Basic model does not solve for the values of the random effects
* Empirical Bayes estimates can be used to predict the values of the random effects (see text)
* Mixed effects models are applicable in a wide variety of data analysis situations
* Unlike Principal Components, Factor Analysis, and other methods we have discussed, there are no “cautions” to using them whenever appropriate