Visualizing Multivariate Data and Models in R

This document is the working outline for the book. It gives some brief descriptions of the topics to be included and the framework to be explained. Currently it is organized in some broad chapters, which might be split into chapters within parts.

## Preface

The main idea here is to explain what can be accomplished with visualizing data from a multivariate lens. I exploit the story of Edwin Abbot’s *Flatland* and give other useful perspectives.

* ONE, TWO, MANY
* Flatland
* EUREKA
* Preliminaries

In here, I also list the main packages used here and data sets for examples.

* R packages: The main substantive packages for multivariate analysis introduced here are heplots, candisc, mvinfluence, VisCollin and matlib. A wide variety of other packages are used for data processing and graphical display. One goal is to present analysis and data display using the tidyverse and ggplot2 framework, but other modern packages are used for informative displays.
* Datasets: I’ve created a separate working document, *Vis-MLM data sets, analyses & plots* listing the main data sets their uses

## Introduction

This chapter attempts to motivate a multivariate approach to data visualization and statistical models.

* why use a multivariate design?
* Linear models: Univariate to multivariate
* visualization is harder
* Problems in understanding and communicating MLM results

## Getting Started

File: 02-getting\_started.qmd

This chapter explains why data visualization is essential in statistical analysis, giving a classical, contrived example and then a real data example. Next, this chapter explains various types of graphical methods useful in data analysis: we distinguish plots of raw data (“data plots”) overlaid with informative graphical summaries showing predicted/fitted values according to some statistical model (“model plots”), and “diagnostic plots” designed to show whether and how the assumptions of the model may be violated. A final section discusses principles for effective graphic display.

* Why plot your data?
  + Anscombe data
  + Davis data
* Data plots
* Model plots
* Diagnostic plots
* Principles of graphic display
  + Criteria for assessing graphs: communication goals
  + Graphic comparison: make important comparisons easy
  + Effect ordering: arrange factors/variable according to what is to be seen
  + visual thinning: remove detail to focus on the main comparisons

## Plots of Multivariate Data

File: 03-multivariate-plots.qmd

This chapter introduces a toolbox of basic graphical methods for visualizing multivariate datasets. It starts with some simple techniques to enhance the basic scatterplot with annotations such as fitted lines, curves and data ellipses to summarize the relation between two variables. These are described as *smoothers*, such as regression lines and nonparametric curves, *stratifiers* to identify subgroups or conditioning in multi-panel displays and *data ellipses* (or more generally bivariate densities).

To visualize more than two variables, we can view all pairs of variables in a scatterplot matrix or shift gears entirely to show multiple variables along a set of parallel axes. As the number of variables increase, we may need to suppress details with stronger summaries (*visual thinning*) for a high-level reconnaissance of our data terrain, as we do by zooming out on a map.

A first draft of this chapter is largely complete.

* Bivariate summaries:
  + Smoothers
  + Stratifiers
  + Data ellipses & bivariate density displays
* Scatterplot matrices
  + Visual thinning
  + Corrgrams
* Generalized pair plots
  + Categorical data: mosaic pairs plots
* Parallel coordinate plots

## Dimension Reduction Techniques: PCA and Biplots

File: 04-pca-biplot.qmd

Beyond a few variables, the limitations of 2D (or even 3D) graphs become quickly apparent. In such cases, it is often profitable to view the data in a low-D space that extracts the most “juice” – the important information within the data. This chapter describes the simplest multivariate juicer, principal components analysis (PCA) and its visualization in the related biplot, which shows the data and the original variables projected onto a space of small dimension.

* PCA, the multivariate juicer
* Biplot, a low-dimensional view

## Overview of Linear Models

In this chapter, I review the standard statistical methods for explaining or predicting a quantitative response using a linear model composed of quantitative and/or categorical predictors.

* Regression
* ANOVA
* ANCOVA
* Discriminant analysis
* Regression trees

## Plots for univariate response models

For a univariate linear model fit using lm(), glm(), and similar functions, the standard plot() method gives basic versions of *diagnostic* plots of residuals and other calculated quantities for assessing possible violations of the model assumptions. Some of these can be considerably enhanced using other packages.

* The “regression quartet”
* Coefficient plots
* Marginal (effect) plots
* Diagnostic plots

## Collinearity and Ridge Regression

This chapter focuses on the problems associated with high correlations among predictors in linear models, which can lead to numerical instability and paradoxical findings that, while a linear model can be highly predictive, few or none of the independent variables appear to be significant.

* Visualizing multicollinearity
  + Collinearity diagnostics
  + Tableplots
  + Collinearity biplots
* Ridge regression – generalized ridge trace plots

## Hotelling’s T^2

Just as the one- and two- sample univariate -test is the gateway drug for understanding analysis of variance, so too Hotelling’s test provides an entry point to multivariate analysis of variance. This simple case provides an entry point to understanding the collection of methods I call the **HE plot framework** for visualizing effects in multivariate linear models, which are a main focus of this book.

* as a generalized -test
* properties
* HE plot and discriminant axis

## Brief Review of the MLM

The general multivariate linear model (MLM) can be understood as a simple extension of the univariate linear model, with the main difference being that there are multiple response variables instead of just one.

This chapter explains the extensions from univariate to multivariate models, focusing on how familiar univariate statistics and methods are translated into their multivariate counterparts. Conceptual and geometric diagrams help to make this understandable.

* ANOVA -> MANOVA
* MRA -> MMRA
* ANCOVA -> MANCOVA
* Repeated measures designs

## Visualizing Multivariate Models

Tests of multivariate models, including multivariate analysis of variance (MANOVA) for group differences and multivariate multiple regression (MMRA) can be easily visualized by plots of a hypothesis (“H”) data ellipse for the fitted values relative to the corresponding plot of the error ellipse (“E”) of the residuals, which I call the *HE plot framework*.

For more than a few response variables, these result can be projected onto a lower-dimensional “canonical discriminant” space providing an even simpler description.

* HE plot framework
* Toy example
* HE plot details
  + Evidence vs. effect scaling
* Canonical discriminant analysis

## Visualizing Equality of Covariance Matrices

Just as univariate ANOVA depends on the assumption that within-group variances are equal, MANOVA tests for group differences depend the analogous assumption that covariance matrices are all the same.

This chapter explains how to visualize this situation using data ellipses for the groups and illustrates a visualization of Box’s M test that is commonly used in this situation.

* Homogeneity of variance in univariate ANOVA
* Homogeneity of variance in MANOVA
* Box’s M test

## Case studies

This is a collection of MANOVA examples, listed here with the sources from other publications I may draw on as case studies or use these in earlier chapters.

### MANOVA Examples

* iris data
* penguins data
* parenting data [from TQMP tutorial]
* diabetes data [from candisc vignette]
* Neuro-Cog & Social-Cog [from TQMP tutorial]

### MMRA Examples

* Rohwer data from [HE-plot-examples vignette & TQMP tutorial]

## Multivariate Influence

This material will most likely go earlier, as an extension of diagnostic plots to the multivariate response setting.

* Univariate influence
* Multivariate influence