Presentation for Final Project - Stats243

AbudulRahman Kreidieh, Aziz Khiyami, Aniket Kesari, Renata Barreto

12/3/2016

Wine Tasting Experiment

- The dataset deals with expert assessors tasting 12 wines from Sauvignon Blanc grapes
  - **Three wine regions**: France, Canada, New Zealand
  - Four wines per region
- 10 experts
  - 9 point scale (1 weakest to 9 strongest)
  - 4 variables
    - Cat Pee
    - Passion Fruit
    - Green Pepper
    - Mineral
  - Additional Variables that the Assessors Added
    - Smoky, Citrus, Tropical, Leafy, Grassy, Flinty, Vegetal, Hay, Melon, Grass, Peach

Wine Tasting Experiment (cont.)

- Other Descriptive Variables
  - Acidity: “Total acidity tells us the concentration of acids present in wine whereas the pH level tells us how intense those acids taste. For example, if you have a wine with 6 g/l total acidity and a pH of 3.2 it will taste more acidic than a wine with 4 g/l total acidity with the same pH level” [@:BDKrEnJ2]
  - pH: Measured ripeness in relation to acidity; low (expected taste: tart, crisp) to high pH (susceptible to bacteria); average pH 3-4, with white wines on the lower end of the spectrum, and red wines on the higher end of that range (Wine Spectator)
  - Alcohol: Alcohol by volume (ABV), measured as the number of milliliters of pure ethanol / 100 milliliters of solution at 20 C; expressed in %
  - Residual Sugar: The sugar that's remaining after fermentation stops (g/L)

Analysis Techniques

- Principle component analysis (PCA): a statistical procedure that uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of linearly uncorrelated variables (principle components)
- Multifactor analysis (MFA) originated with French statisticians Brigitte Escofier and Jerome Pages $\sim$ 1980
- Initially, MFA was a response or generalization of PCA
- It has a number of goals:
  - analyze several data sets measured on the same observation
  - provide a set of common factor scores (compromise factor scores)
  - project each of the original data sets onto the compromise to analyze communalities and discrepancies

PCA vs. MFA

- MFA is part of the PCA family and hence its main analytical tool is SVD and GSVD
- In MFA, like PCA, the importance of a dimension (ie principle component) is reflected by its eigenvalue which indicates how much of the total inertia (variance) of the data is explained by this component

Multifactor Analysis: Steps and Outcomes

- Steps of MFA:

1. `A PCA of each data table is performed and the first sin`
2. `A grand matrix is created by concatenating all the data`
3. `The observation partial factor scores for each table ar`

- Dervied Outputs of MFA
- Summarized information about the obtained **eigenvalues** ("explained inertia")

Contributions

- **Contributions**, a series of descriptive statistics that allow us to interpret how the following contribute to the extracted dimensions:
  1. observations
  $$ctr_{i,l} = \frac{m_i * f_{i,l}^2}{\lambda_l}$$
  2. the variables
  $$ctr_{j,l} = a_j * q_{j,l}^2$$
  3. the tables
  $$ctr_{k,l} = \sum_j^{J_k} ctr_{j,l}$$

Sample Code + Plots

- The basic idea behind all three of these contribution measures is to weight each assessor's contribution to the overall score
  - Contributions range from [0,1]
  - Contributions sum to 1
  - The larger the contribution, the more weight it has in explaining the overall score
- Essentially, the MFA method takes a series of blocked observations, standardizes them, and reports them in a simple vectorized weight format
- Below, we'll provide some snippets from the code to carry out these steps. Full code, with comments, is available in the package

Step One: PCA of Each Data Table

```
# Step 1: PCA of Each Data Table
F_partial = list() # Empty list for partial factor
a = c() # Empty list for alphas
K = length(sets) # Empty list of length sets for Ks
J = c() # Empty vector for Js
index = 1 # Start index at 1
for (i in 1:length(sets)) {
# Break up data into each assessor
Xi = dat[,index:(index+length(sets[[i]])-1)]
J = c(J, length(sets[[i]]))
1
# Compute SVD using the "svd" function in the base package
SVD = svd(Xi) # Save results of svd on the table of assesso
U = SVD$u # Pulls a matrix whose columns contain the left s
D = diag(SVD$d) # Pulls a vector containing the singular va
V = SVD$v # Pulls a matrix whose columns contain the right
# alpha weights
alpha_1 = D[1,1]^-2 # Weight is equal to the reciprocal squ
```

Step Two: Grand Matrix for GSVD

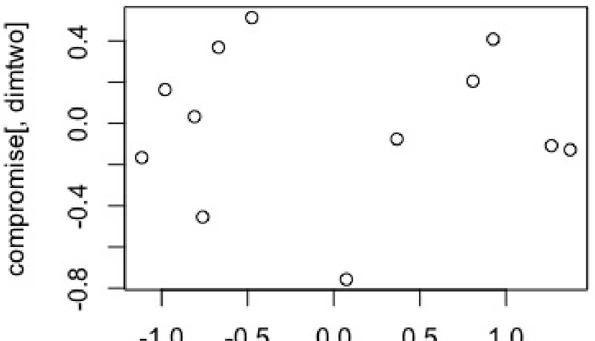
```
= rep(1/dim(dat)[1], dim(dat)[1]) # Table of dimensions
# Compute GSVD
GSVD = svd(diag(m^(1/2)) %*% dat %*% diag(a^(1/2))) # Apply
```

Step Three: Partial Factor Scores

```
# Partial Factor Scores (step 2)
index = 1
for (i in 1:length(sets)) {
F_partial[[i]] = F_partial[[i]] %*% t(Q)[index:(index+lengt
index = index + length(sets[[i]])
```

Sample Plot: Common Factor Score

- After running the package and the associated auxillary print and plot functions, we can plot individual elements from the mfa object
- Below is an example of a plot of the 1st and 2nd columns from the compromise score table



RV Coefficient + Lg Coefficient

- Used to evaluate the similarity between two tables
  - In other words, this can be interpreted as a noncentered squared coefficient of correlation between two matrices; it reflects the amount of variance shared by two matrices
  - Varies between 0 – 1
- Lg Coefficient
  - Reflects MFA normalization and takes positive values

Bootstrapping

- Used to estimate the stability of the compromise factor scores
- The stability of the descriptive statistics (i.e. contributions of observations, variables, and tables to a dimension) can be determined with boostrapping, which is a cross validation technique whose results can be used to select the relevant elements of a dimension

Shiny App

- We implemented the package in a Shiny app interface
- The app provides a dropdown menu in which the user can specify whether they would like to plot the eigenvalues, common factor scores, partial factor scores, or factor loadings
  - If the user chooses one of the latter three options, they can further specify which columns to plot

Bibliography

- Vinny, D. (2016). "Ask Dr. Vinny." from http://www.winespectator.com/drvinny/show/id/5035.
- Abdi, H., et al. (2013). "Multiple factor analysis: principal component analysis for multitable and multiblock data sets." Wiley Interdisciplinary reviews: computational statistics 5(2): 149-179.
- (2015). "Understanding Acidity in Wine." from http://winefolly.com/review/understanding-acidity-in-wine/.