Display of Multivariate Data

Math 257 Dr. Andrea Gottlieb

San Jose State University
Milica Miskovic, Shuai Li, Yuting Tao

May 8, 2019

Chernoff faces

Named after their inventor Herman Chernoff (1973). Representing p-dimensional observations as a two-dimensional face whose characteristics (face shape, mouth curvature, nose length, eye size, and so forth) are determined by the measurements on the p<18 variables.

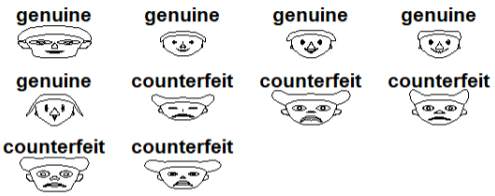

Figure 1: *Chernoff faces with 6 parameters*

Advantages

Most useful for verifying (1) an initial grouping suggested by subject-matter knowledge and intuition or (2) final groupings produced by clustering algorithms. More applicable to global applications such as clustering and identification of trends. (3)They look funny.

Drawbacks

Faces requires training in application and interpretation. Potentially inappropriate for certain types of social data, not optimal for large amount of observations or large number of varables.

code

```
install.packages("aplpack")
crime <- read.csv("http://datasets.flowingdata.com/
crimeRatesByState-formatted.csv")
library(aplpack)
faces(crime[,2:8])
?faces #to see which variable is represented
by which feature
```

Star plots

In two-D, we can construct circles of a fixed radius with p equally spaced rays from the center of the circle. Each observation is represented as a star-shaped figure with one ray for each variable. The length of each ray is proportional to the value of its corresponding variable. The open ends of the rays are usually connected with lines.

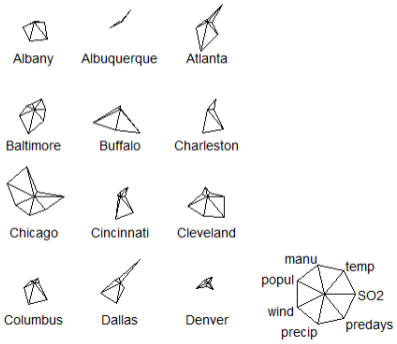

Figure 2: *Star plot*

Advantages

Quick and intuitive, easy to spot outliers. Also known as web chart, spider chart, cobweb chart, irregular polygon, polar chart, or Kiviat diagram.

Drawbacks

As the number of rays increases, it becomes more difficult to separate them. The number of distinguishable arrays may be increased by adding retinal visual properties e.g. hue, luminance, width, etc. Observations have to be standardized so that the center of the circle represents the smallest observation within the data set.

code

```
data(mtcars)
stars(mtcars[, 1:7], len = 0.8, key.loc = c(12, 2),
      main = "Motor Trend Cars", full = FALSE)
```

Parallel coordinates

Parallel coordinates plots represent multidimensional data using lines. A vertical line represents each variable each line is a collection of points placed on each axis, that have all been connected together. Overlapping of line segments occurs when many data records have the same or similar values or the number of data records is large relative to the display.

The order the axes are arranged in can impact we understand the data. One reason for this is that the relationships between adjacent variables are easier to perceive than for non-adjacent variables. So re-ordering the axes can help in discovering patterns or correlations across variables.

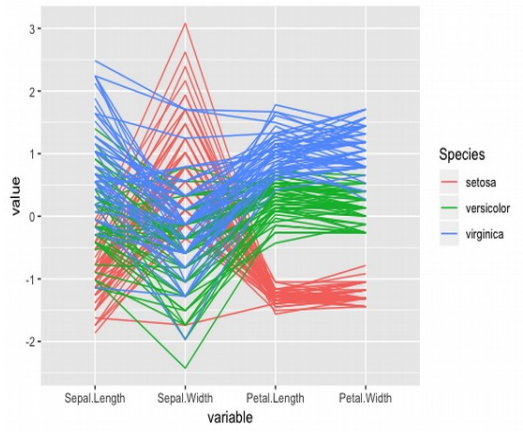

Figure 3: *Parallel coordinates*

Advantages

Distinguish which variables dominate a single observation, helpful for detecting outliers.

Drawbacks

Ordering of variables has influence on visualization and interpretation. Neighboring relationships can be lost in visualization and others can be more emphasized

code

```
library(GGally)
data(iris)
p <- ggparcoord(data = iris, columns = 1:4,
   groupColumn = 5, order = "anyClass",
   showPoints = TRUE, title = "Parallel Coordinate Plot
   for the Iris Data", alphaLines = 0.3)
p_(p)
```

Violin plots

Violin plots are similar to box plots, except that they also show the number of points at a particular value by the width of shape. They can also include the marker of median and a box for a interquartile range. Virginica has highest median value in petal length, petal width and sepal length when compared against Versicolor and Setosa. However, Setosa has the highest sepal width median value. We can also see significant difference between SetosaâĂŹs sepal lenght and width against its petal length and width. That difference is smaller in Versicolor and Virginica. The violin plot also indicates that the weight of the Virginica sepal width and petal width are highly concentrated around the median

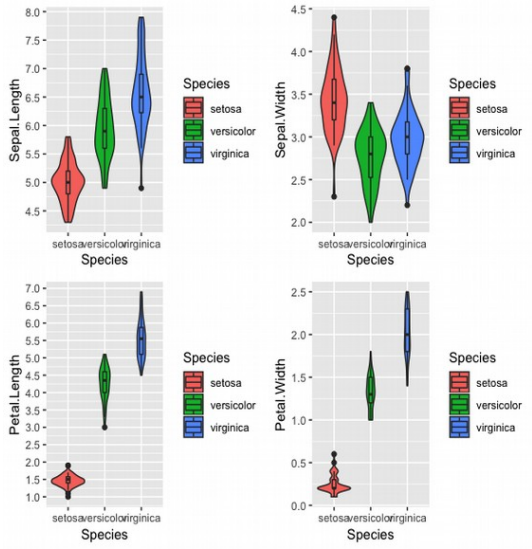

Figure 4: *Violin plots*

Advantages

Easy to follow ind interpret.

Drawbacks

Multiple univariate plots.

code

```
library(gridExtra)
library(ggplot2)
library(tidyverse)
library(stringr)
gather(iris, Var, value, -Species) %>% ggplot(aes(Var, value))+
  geom_violin(aes(fill = Species))+
  facet_grid(~Species)+
  theme(axis.text.x = element_text(angle = 90, vjust = .5))+
  labs(x = "Measurements", y = "Length in cm", title = "Violin
  geom_boxplot(width=0.1)
```

Andrews plots

Introduced by Andrews (1972). Andrews plot displays high dimensional data point as a parametric curve in terms of new parameter **t**, developed in a Fourier series (or other) .A point $y = (y_1, y_2, \ldots, y_p)$, is displayed as the curve of the function of parameter t with those values $y_1, y_2, \ldots, y_p$ as coefficients. $f_y(t) = \frac{y_1}{\sqrt{2}} + y_2 sin(t) + y_3 cos(t) + y_4 sin(2t) + y_5 sin(2t) + \ldots$

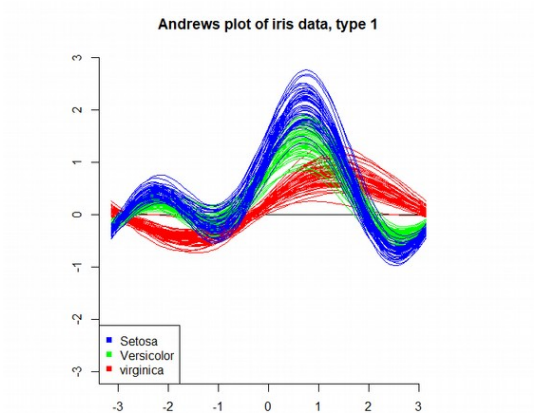

Figure 5: *Andrews plots*

Advantages

Perfect for mean values detection, outlier detection and identifying clusters (the points which are closer to each other are the curves nearer to each other). Principal component count.

Drawbacks

The shape depends on the order of variables in Fourier series. The shape depends on the function we use. Then $t = 0$, all even numbered terms in the Andrews function simultaneously vanish.

code

```
library(andrews)
andrews(iris, ymax=3, type = 1, clr = 5,
 main = "Andrews plot of iris data, type 1")
legend("topright", legend = c("Setosa",
 "Versicolor", "virginica"), col = c("blue",
 "green", "red"), pch = 15)
```

Radial visualization

Each observation is represented with point on 2d radial diagram. The value of the variable is the force of the theorethical spring that pulls the point towards the outer edge. If the values of all variables are similar, point will be in center. If variable is influential, it will pool the point towards its edge.

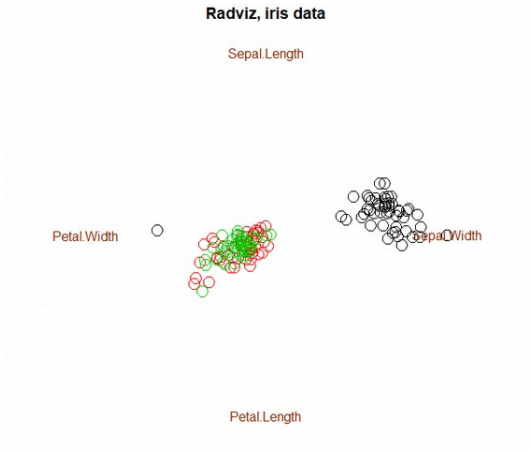

Figure 6: *Radial visualization*

Advantages

Cluster detection. Excellent with big data sets with lots of observations. It is easy to detect and select an influential variable.

Drawbacks

It is not straightforward to implement in R. The order of the variables is important since different arrangements of variables give different display of the data, therefore it requires backgound knowelade for selection of the most informative graph (some packages do it automatically). The data should be standardized so that every variable pulls the point according to the value, not the scale. Large number of variables can compromise the interpretability.

code

```
library(Radviz)
iris_norm <- apply(iris[,1:4],2,do.L,
 fun=function(x) quantile(x,c(0.025,0.975)))
ct.S <- make.S(colnames(iris[,1:4]))
ct.sim <- cosine(iris_norm)
ct.rv <- do.radviz(iris_norm,ct.S)
plot(ct.rv, point.shape=1, point.size = 2,
 point.color=as.integer(iris$Species),
main = "Radviz, iris data")
```

References

[Chernoff faces, star plots] Johnson and Wichern,(2008). Applied Multivariate Statistical Analysis, 6nd ed.

[Chernoff faces, star plots] Harold W. Gugel,(2008). Stars and Faces, Procedure for Constructing Graphical Profiles of Multivariate Data.

url = hhttp://www.sascommunity.org/sugi/SUGI85/Sugi-10-46

[Violin plots] Hintze Jerry and Nelson Ray,(1998). Violin Plots: A Box-Plot Dencity Trace Synergism

url = http://www.stat.cmu.edu/ rnugent/ PCMI2016/papers/ViolinPlots.pdf

[Paralel plots] Rida E. A. Moustafa and Edward J. Wegman. On Some Generalizations ofParallel Coordinate Plots

url = https://tinyurl.com/yyllnfxx

[Radial plots] Patrick Hoffman and Georges Grinstein. Visualizations for High Dimensional Data Mining,Table Visualizations.

url = https://tinyurl.com/yxvb84ut
url = https://cran.r-project.org/web/packages/Radviz/Radviz.pdf
url = https://tinyurl.com/y2dgq6uk

[Andrews plots] Patrick Hoffman and Georges Grinstein. Visualizations for High Dimensional Data Mining,Table Visualizations.
url = https://tinyurl.com/y2dgq6uk
url = https://cran.r-project.org/web/packages/andrews/andrews.pdf