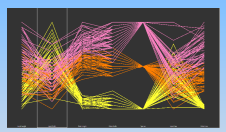
Parallel Coordinates



Multivariate data higher than 3D are hard to visualize in cartesian coordinates.

Many different methods have been suggested for visualizing multivariate (dimensionality > 3) data (e.g. Principal components).

Parallel Coordinates of Inselberg (1, 2) is one alternative that is limited only by the size of the monitor screen.

(1) Don't panic ... just do it in parallel! Al Inselberg, COMPUTATION STAT 14: (1) 53-77 1999
(2) Visual data mining with parallel coordinates Al Inselberg, COMPUTATION STAT 13: (1) 47-63 1998

Parallel Coordinates in R

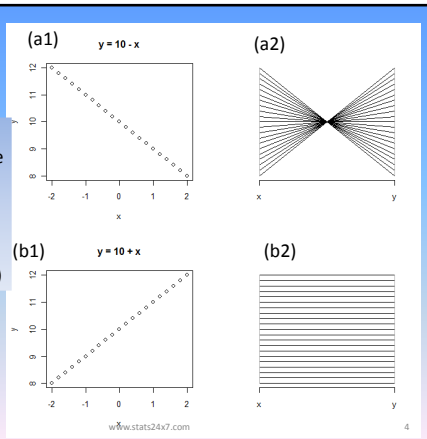
1) Install package MASS in R.
2) Load MASS.
3)
```
nf <- layout(matrix(c(1,2,3,4),2,2,byrow=TRUE), TRUE)
# shows the graphics layout
 layout.show(nf)
```

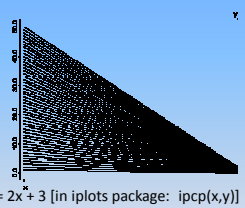
4)
```
a <- read.csv("K:/TEACH/DataMining_Fall2009/Data/line3.csv",header=TRUE)
attach(a)
plot(x,y, main = "y = 10 - x")
parcoord(a, col = 1, lty = 1, var.label=FALSE)

b <- read.csv("K:/TEACH/DataMining_Fall2009/Data/line4.csv",header=TRUE)
attach(b)
plot(x,y, main = "y = 10 + x")
parcoord(b, col = 1, lty = 1, var.label=FALSE)
```


Plots of
(a)Decreasing line
(b)Increasing line

in Cartesian (1) and parallel coordinates (2)





Parallel plot of y = 2x + 3 [in iplots package: ipcp(x,y)]

Parallel Coordinates in R

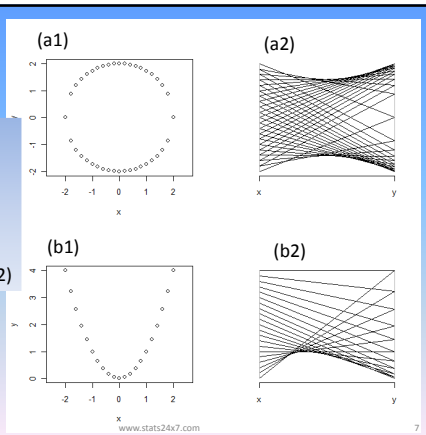
1) Install package MASS in R.
2) Load MASS.
3)
```
nf <- layout(matrix(c(1,2,3,4),2,2,byrow=TRUE), TRUE)
# shows the graphics layout
 layout.show(nf)
```

4)
```
c <- read.csv("K:/TEACH/DataMining_Fall2009/Data/circle.csv",header=TRUE)
attach(c)
plot(x,y, asp=1)
parcoord(c, col = 1, lty = 1, var.label=FALSE)

d <- read.csv("K:/TEACH/DataMining_Fall2009/Data/parabola.csv",header=TRUE)
attach(d)
plot(x,y, asp=1)
parcoord(d, col = 1, lty = 1, var.label=FALSE)
```

Plots of
(a)Circle
(b)Parabola

in Cartesian (1)
and parallel
coordinates (2)



R has several packages for parallel coordinates. An interactive R package is iplot.

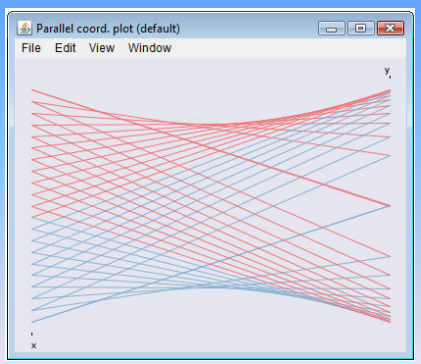
Install and load iplots.

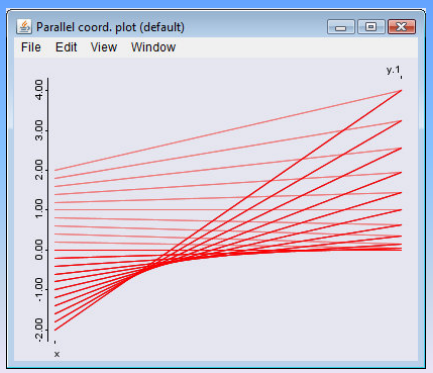
Read the data file circle.csv

```
c <-
    read.csv("K:/TEACH/DataMining_Fall2009/Data/circle.csv",header
    =TRUE)
attach(c)

Ipcp(c)
```

Repeat for the file parabola.csv.





Exercise to do in class:

1. Fit $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ to data in the file anyregression.csv.

Does the fitted model provide good fit to the data?

2. Repeat for $Y_1 = \ln(Y)$ and $Y_2 = \sqrt{Y}$.


```
Call:
lm(formula = Y ~ X1 + X2 + X3)

Residuals:
   Min     1Q  Median     3Q    Max
-47.283 -14.245   1.012  13.757  59.142

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  97.8002     2.3164  42.222  < 2e-16 ***
X1            1.3774     0.3280   4.199 4.64e-05 ***
X2            0.7369     0.3518   2.095  0.03791 *
X3            0.9420     0.3560   2.646  0.00903 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.06 on 146 degrees of freedom
Multiple R-squared: 0.9676,    Adjusted R-squared: 0.967
F-statistic:  1455 on 3 and 146 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = lnY ~ X1 + X2 + X3)

Residuals:
     Min       1Q   Median       3Q      Max
-0.428783 -0.079935  0.005036  0.090634  0.388577

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.703160   0.016245 289.506  < 2e-16 ***
X1          0.007665   0.002300   3.332  0.00109 **
X2          0.003177   0.002467   1.288  0.19989
X3          0.002955   0.002497   1.184  0.23846
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1407 on 146 degrees of freedom
Multiple R-squared: 0.9254,     Adjusted R-squared: 0.9239
F-statistic: 604.1 on 3 and 146 DF,  p-value: < 2.2e-16
```



```
Call:
lm(formula = sqrtY ~ X1 + X2 + X3)

Residuals:
    Min      1Q  Median      3Q     Max
-2.22719 -0.53358  0.02841  0.52334  2.38650

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.27832    0.09346 109.981  < 2e-16 ***
X1           0.05064    0.01323   3.827 0.000192 ***
X2           0.02401    0.01419   1.692 0.092790 .
X3           0.02672    0.01436   1.861 0.064811 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8093 on 146 degrees of freedom
Multiple R-squared: 0.9529,     Adjusted R-squared: 0.9519
F-statistic: 984.3 on 3 and 146 DF,  p-value: < 2.2e-16
```


- Why are all these different models providing very good fit to this data set?
- Install package 'iplots' in R, load it, then draw parallel coordinate plot of the data using the command

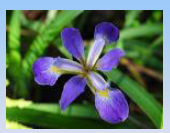
  lpcp(data)

Parallel Coordinates Plot of the Iris Data Set

- Iris data is perhaps the best known database in the pattern recognition/clustering literature. Fisher's paper is a classic in the field and gets cited to this day. The data set contains 3 classes of 50 cases each, each class referring to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.
- Attribute to be Predicted attribute = class of iris plant.

1. Fisher,R.A. "The use of multiple measurements in taxonomic problems" Annual Eugenics, 7, Part II, 179-188 (1936); also in "Contributions to Mathematical Statistics" (John Wiley, NY, 1950).

Iris versicolor

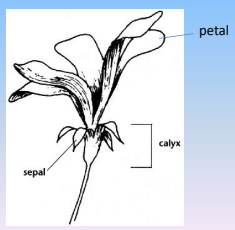


Iris virginica

Iris setosa

Iris Data Set: Length and width of sepal and petal for three northern american species of iris.

The data on *iris setosa canadensis* and *iris versicolor* has been used by R. A. Fisher to illustrate linear discriminance analysis. The data on *iris virginica* have been added as an extension.



Open the data set iris.csv in R.

Calculate Sepal.Area and Petal.Area.

Calculate Sepal.Ratio and Petal.Ratio (ratio = length/width).

Plot (label points by species)

Petal.Length vs Petal.Width

Sepal.Length vs Sepal.Width

Petal.Area vs Sepal.Area

Petal.Ratio vs Sepal.Ratio

What can you say from these graphs?