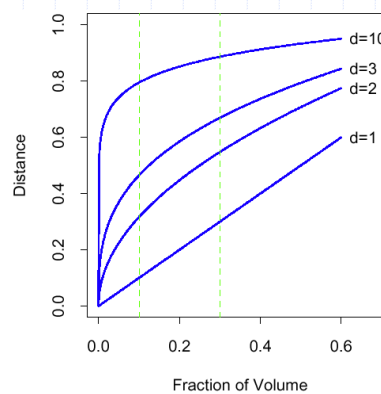
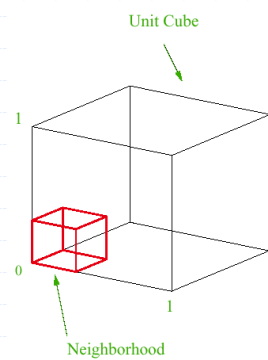


Feature selection and discrimination in high dimension

Randall Julian
Lilly Research Laboratories

Local methods in high dimensions

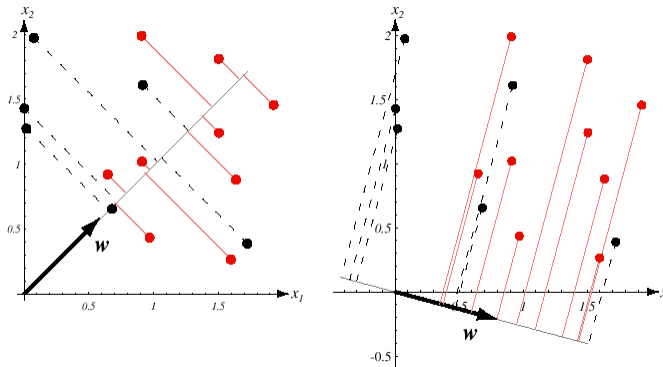
◆ "Curse of dimensionality"



In 10 dimensions we need to cover 80% of the range of each coordinate to capture 10% of the data.

[1]

Projection to a new axis:



[2]

Linear Discriminant Functions

sample mean vectors

$$\bar{\mathbf{x}}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} \mathbf{x}_{ji} \quad , \quad j = 1, 2, \dots$$

Sample covariance matrix

$$\mathbf{S}_j = \frac{1}{N_j - 1} \sum_{i=1}^{N_j} (\mathbf{x}_{ji} - \bar{\mathbf{x}}_j) (\mathbf{x}_{ji} - \bar{\mathbf{x}}_j)^T \quad , \quad j = 1, 2, \dots$$

Pooled sample covariance matrix

$$\mathbf{S} = \frac{1}{N_1 + N_2 - 2} [(N_1 - 1)\mathbf{S}_1 + (N_2 - 1)\mathbf{S}_2]$$

[3]

Standard Distances to discriminants

multivariate standard distance

$$D(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2) = \max \frac{\mathbf{a}^T (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)}{(\mathbf{a}^T \mathbf{S} \mathbf{a})^{1/2}}$$

multivariate standard distance (nonsingular S)

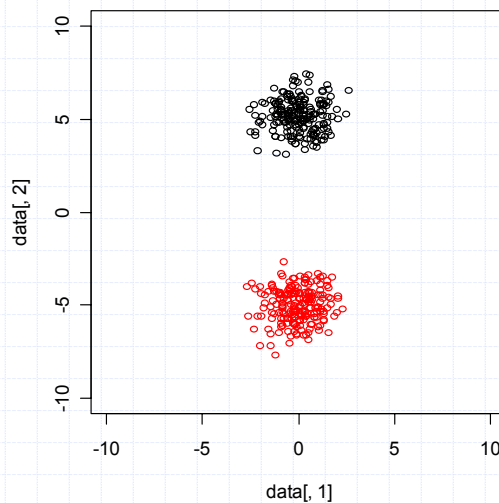
$$D(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2) = \left[(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{S}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \right]^{1/2}$$

vector of coefficients of the linear discriminant function:

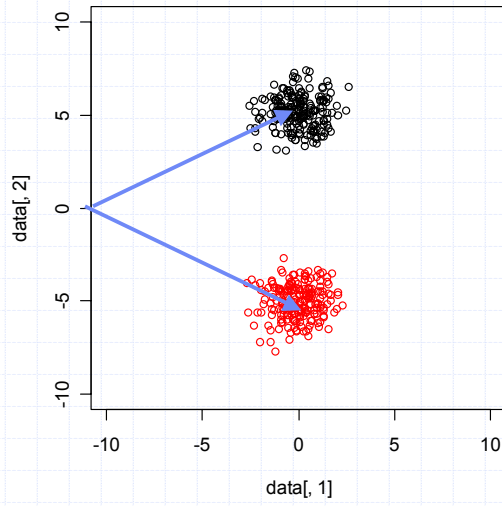
$$\mathbf{b} = \mathbf{S}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

[3]

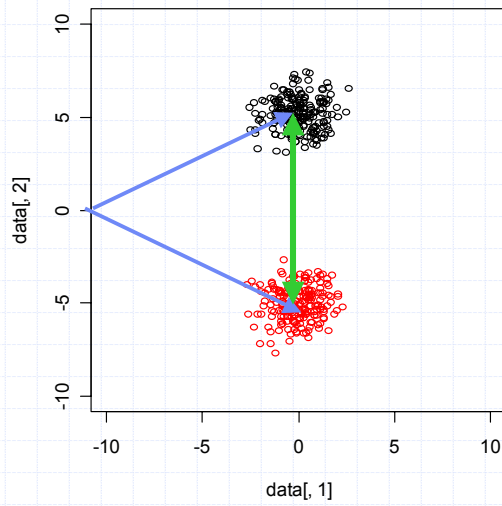
Simple 2D Example



Simple 2D Example



Simple 2D Example



In R:

```
library(MASS)

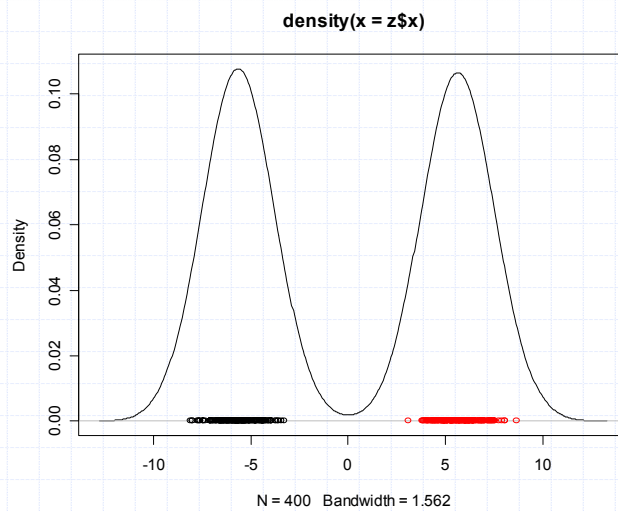
g <- lda( y ~ x1 + x2 , data = data)

Z <- predict(g, Xcon)
zp <- Z$post[,1] - Z$post[,2]

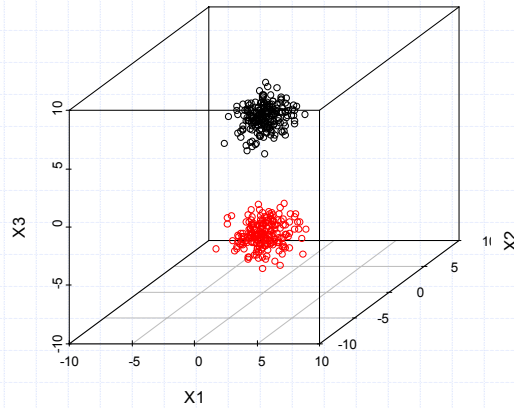
contour(x,y,matrix(zp,length(x),length(y)), add=T, levels=0, labex = 0)
```

[4]

Density on new axis

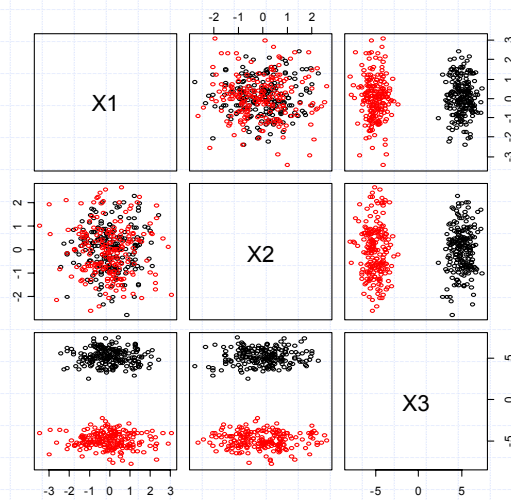


3D Example



package: scatterplot3d (add package from CRAN)
library(scatterplot3d)

Viewed as a multiplot



plot(x1,x2,x3, col=y)

References

- [1] "The Elements of Statistical Learning: Data Mining, Inference and Prediction", Hastie, Tibshirani and Friedman, Springer-Verlag, (2001).
- [2] "Pattern Classification", Duda, Hart, Stork, John Wiley Sons, (2001).
- [3] "A First Course in Multivariate Statistics", B. Flury, Springer, 1997
- [4] "Modern Applied Statistics in S", W.N. Venables, B.D. Ripley - 4th Ed, Springer-Verlag, 2002