PCA etc.

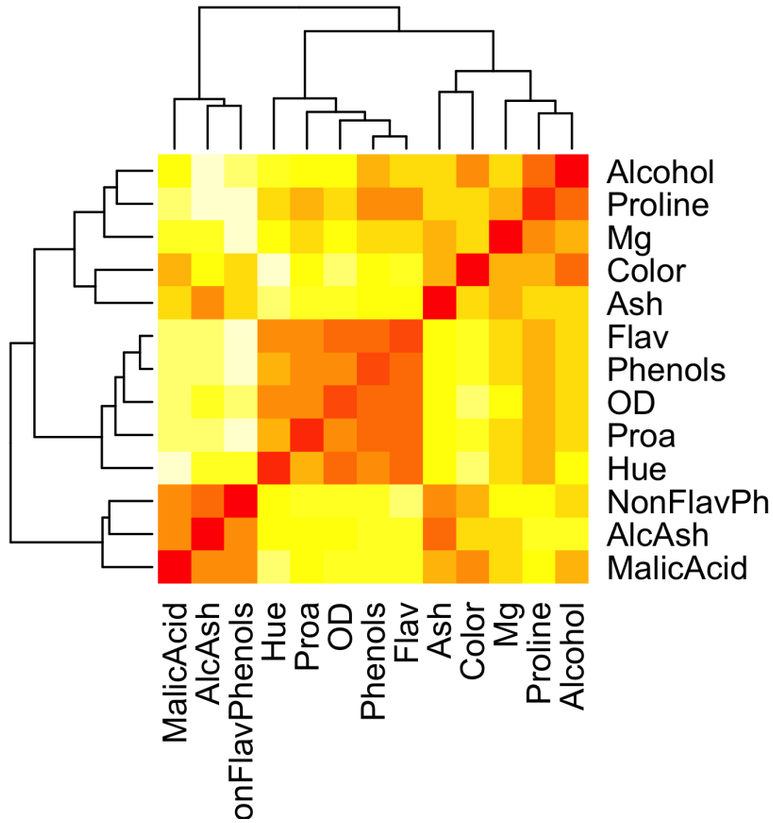
Wolfgang Huber

CSAMA 2015

```r
require(ggbiplot)
data(wine)
wine[1:3,1:7]
```

```
##   Alcohol MalicAcid  Ash AlcAsh  Mg Phenols Flav
## 1    14.2      1.71 2.43   15.6 127    2.80 3.06
## 2    13.2      1.78 2.14   11.2 100    2.65 2.76
## 3    13.2      2.36 2.67   18.6 101    2.80 3.24
```

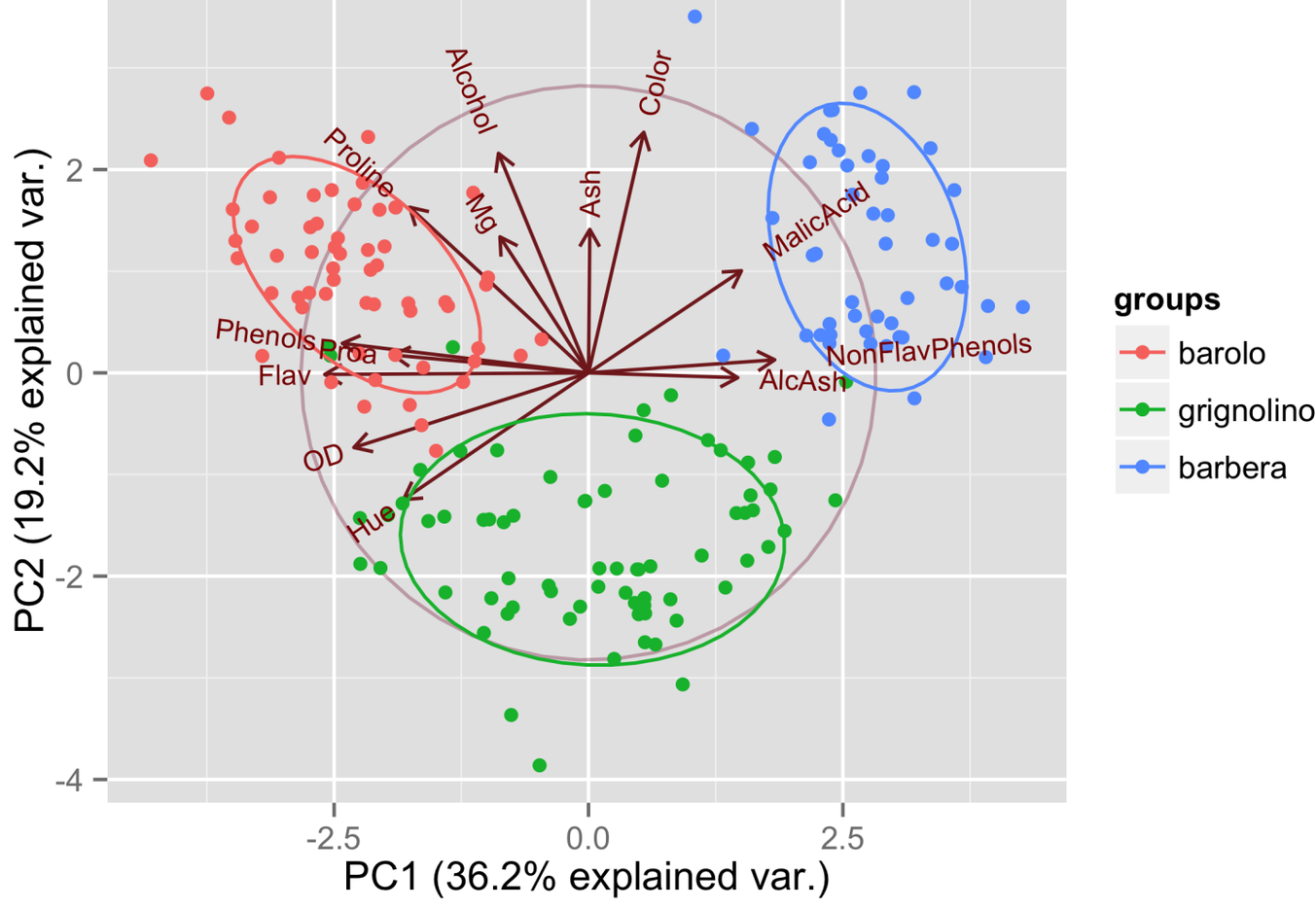
```r
heatmap(1-cor(wine))
```



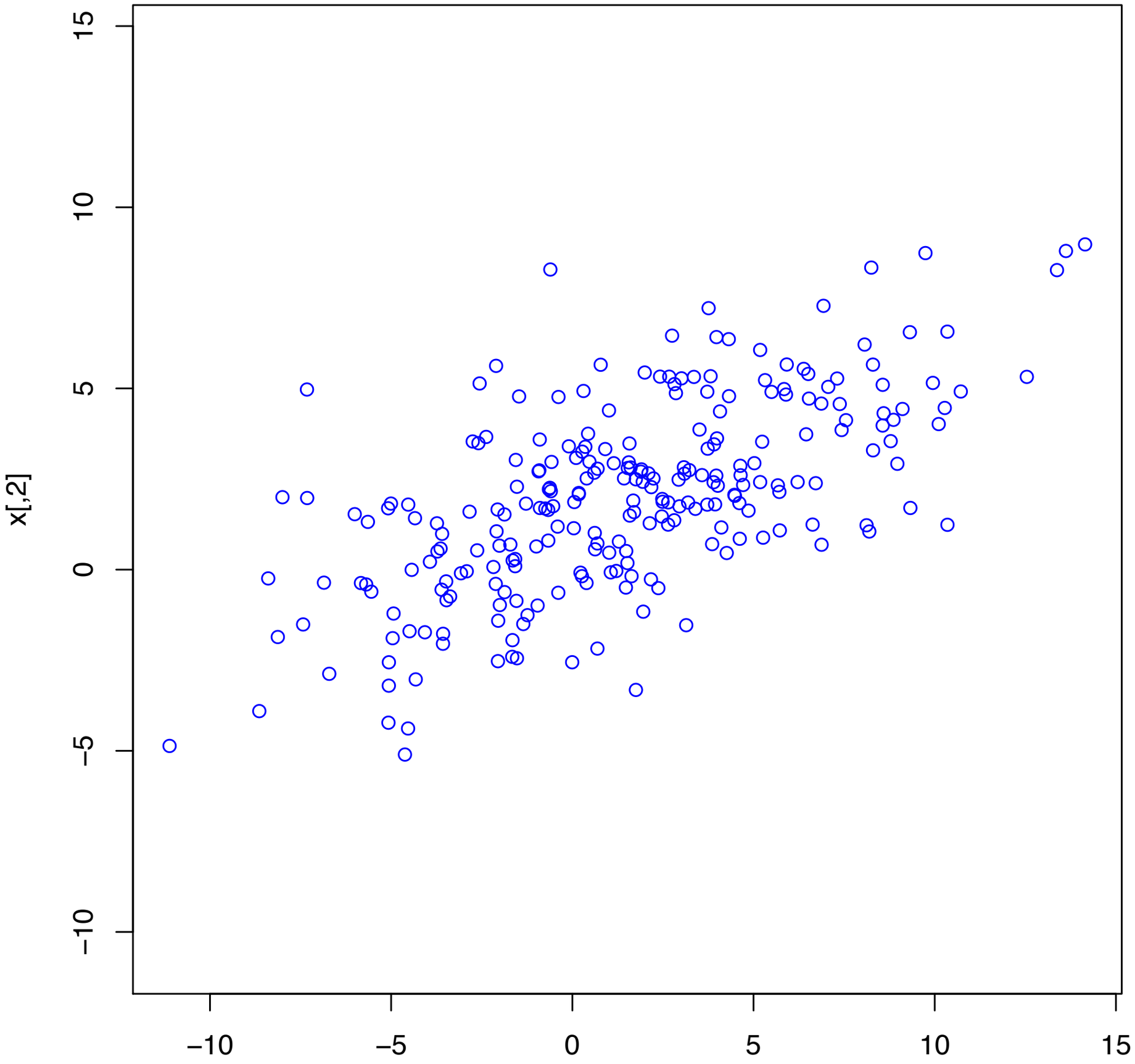

```r
wine.pca <- prcomp(wine, scale. = TRUE)
table(wine.class)
```

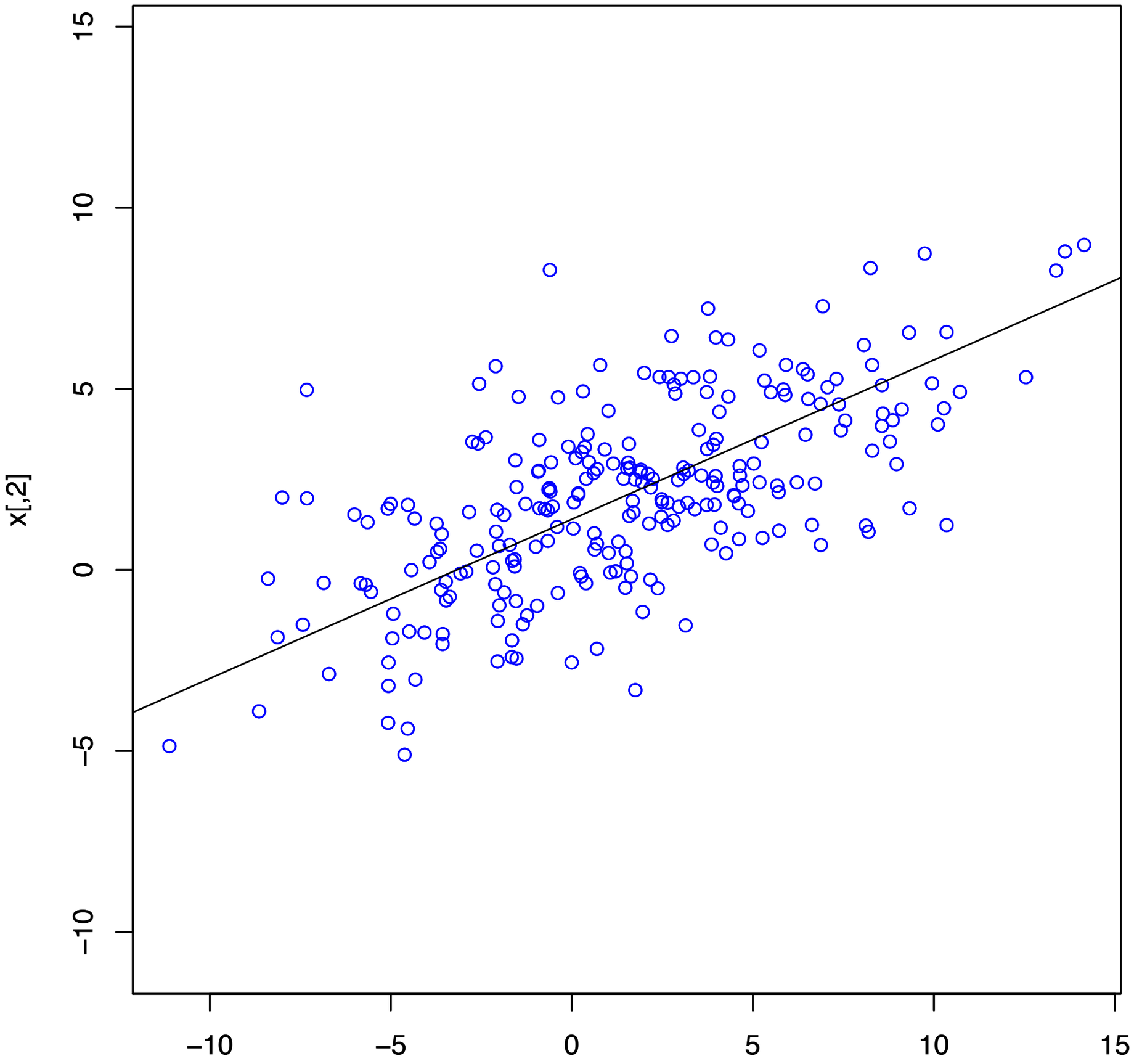
```
## wine.class
##     barolo grignolino    barbera
##         59         71         48
```

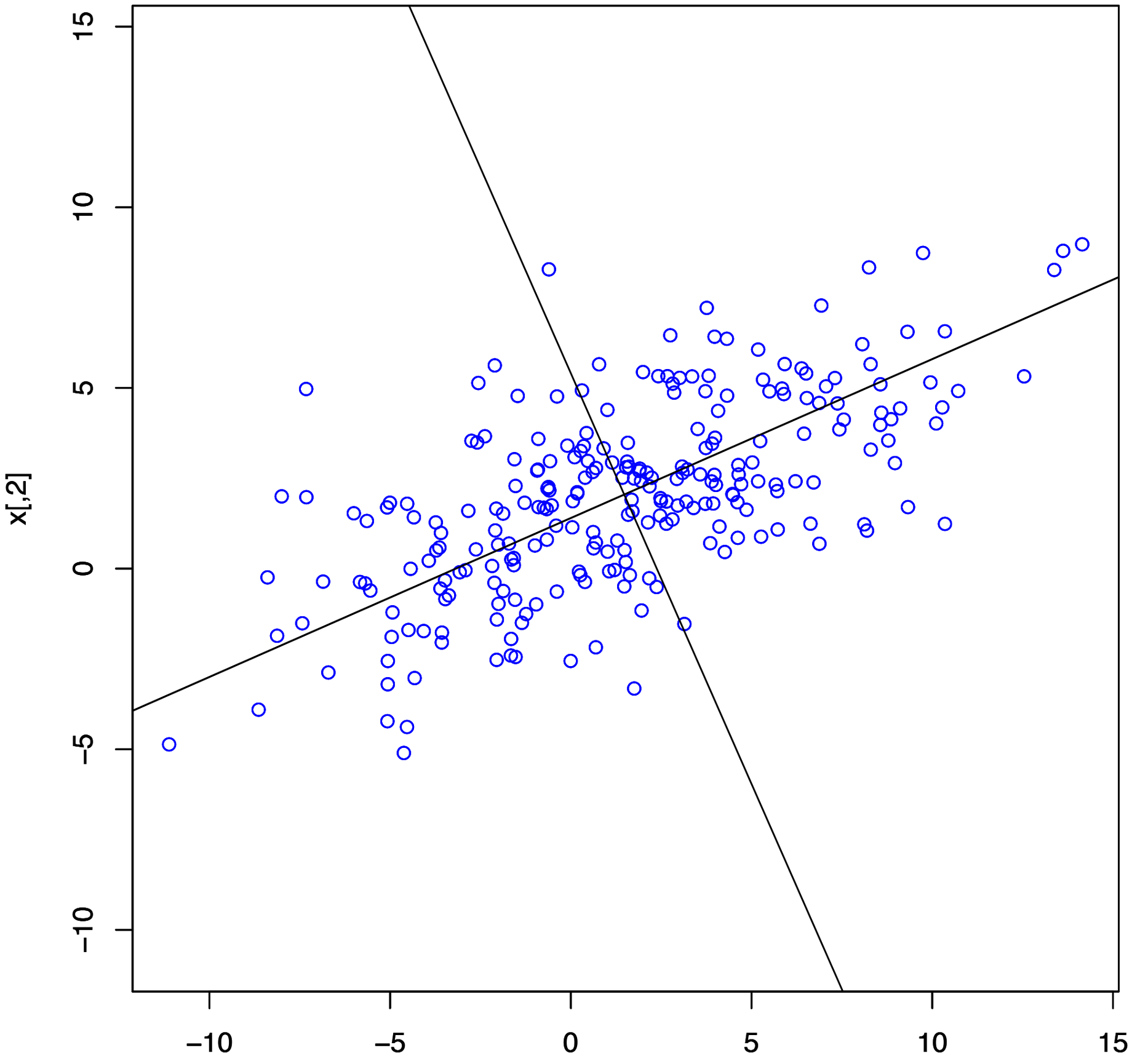
```r
pp=ggbiplot(wine.pca, obs.scale = 1, var.scale = 1,scale=1,
  groups = wine.class, ellipse = TRUE, circle = TRUE)
pp
```



Susan Holmes


x[,2]
15
10
5
0
−5
−10
−10
−5
0
5
10
15
x[,1]


x[,2]
15
10
5
0
−5
−10
−10
−5
0
5
10
15
x[,1]


x[,2]
15
10
5
0
−5
−10
−10
−5
0
5
10
15
x[,1]

Principal Component Analysis

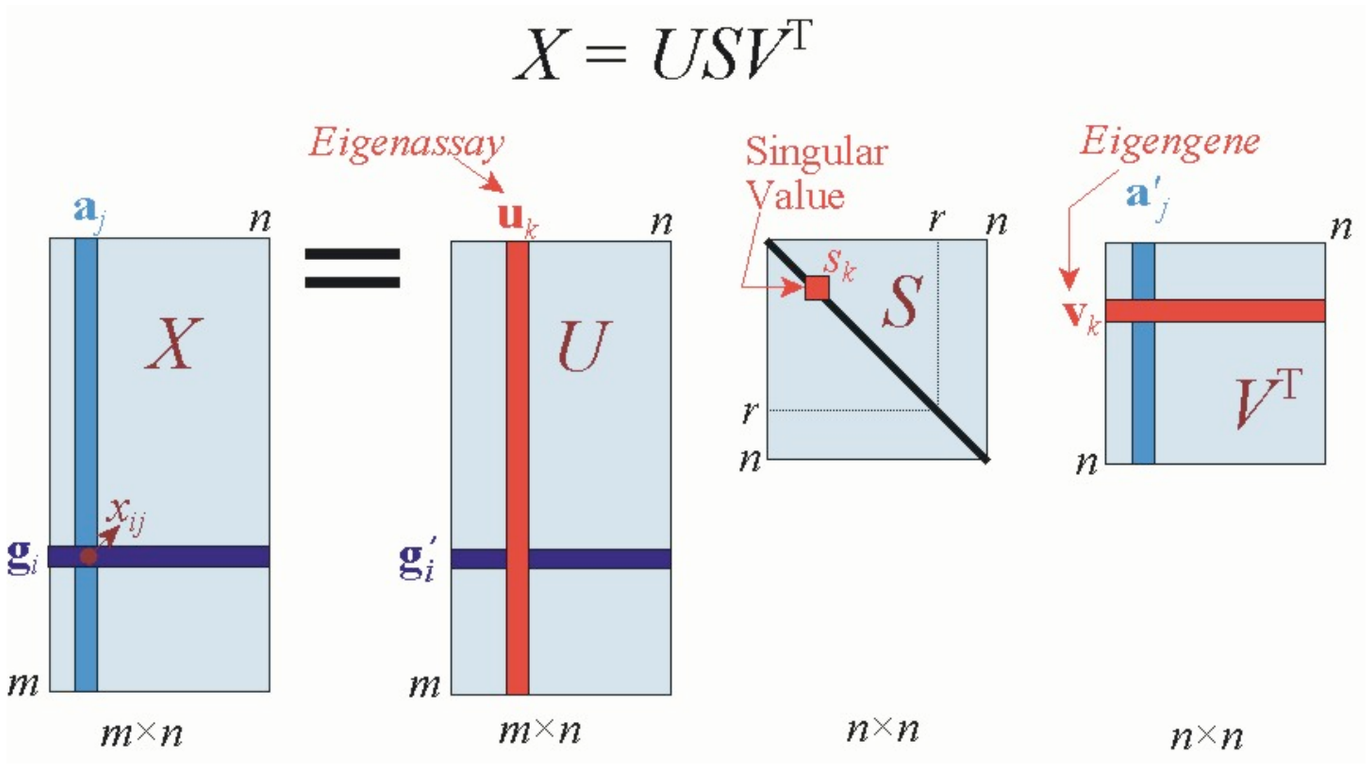
Data points $x_i \in \mathbb{R}^n$

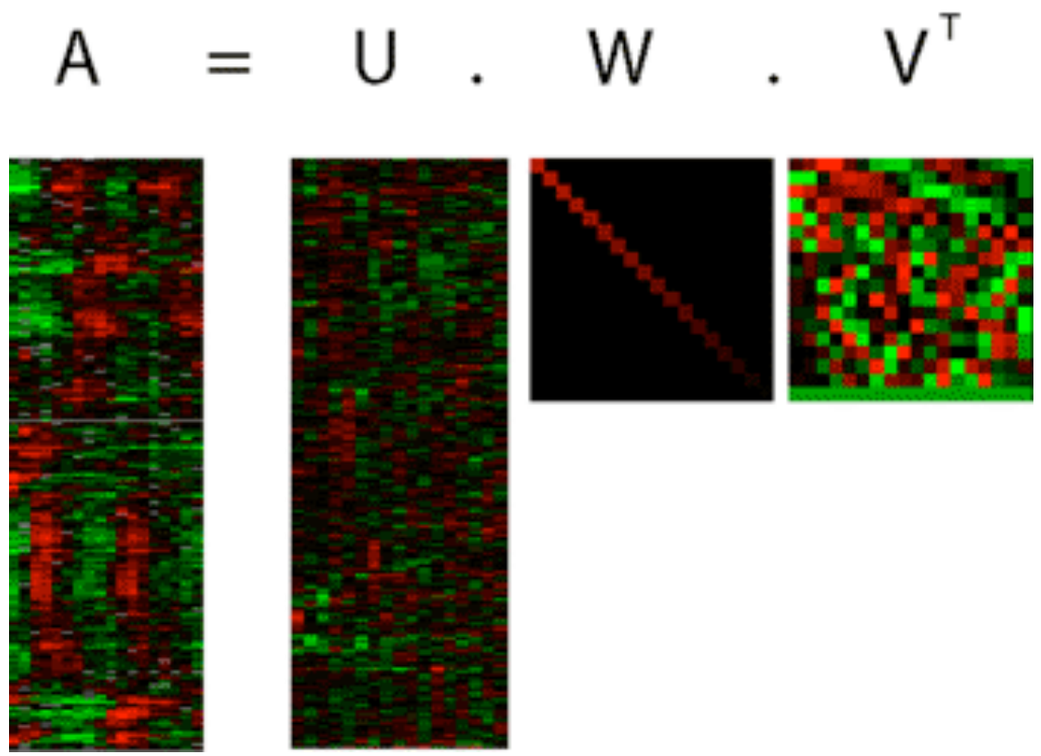
Linear projection $P: \mathbb{R}^n \to \mathbb{R}^k \quad (n \le k)$

$$\sum_i (x_i - P(x_i))^2 \quad \to \quad \min$$

$$\text{Cov}\, P(x_i))^2 \quad \to \quad \max$$

How is the Principal Component Analysis computed?





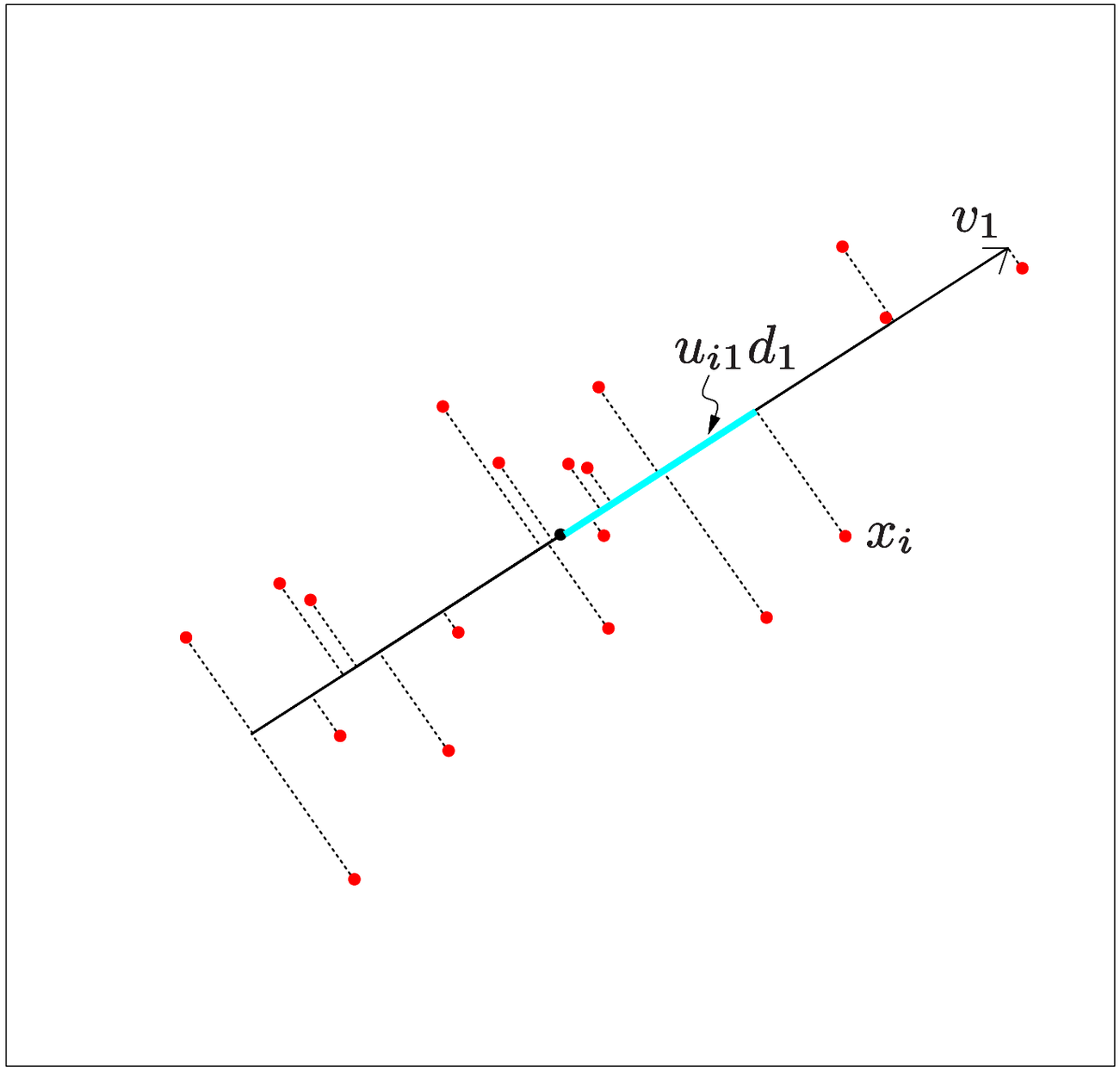


FIGURE 14.20. *The first linear principal component of a set of data. The line minimizes the total squared distance from each point to its orthogonal projection onto the line.*

Hastie, Tibshirani, Friedman

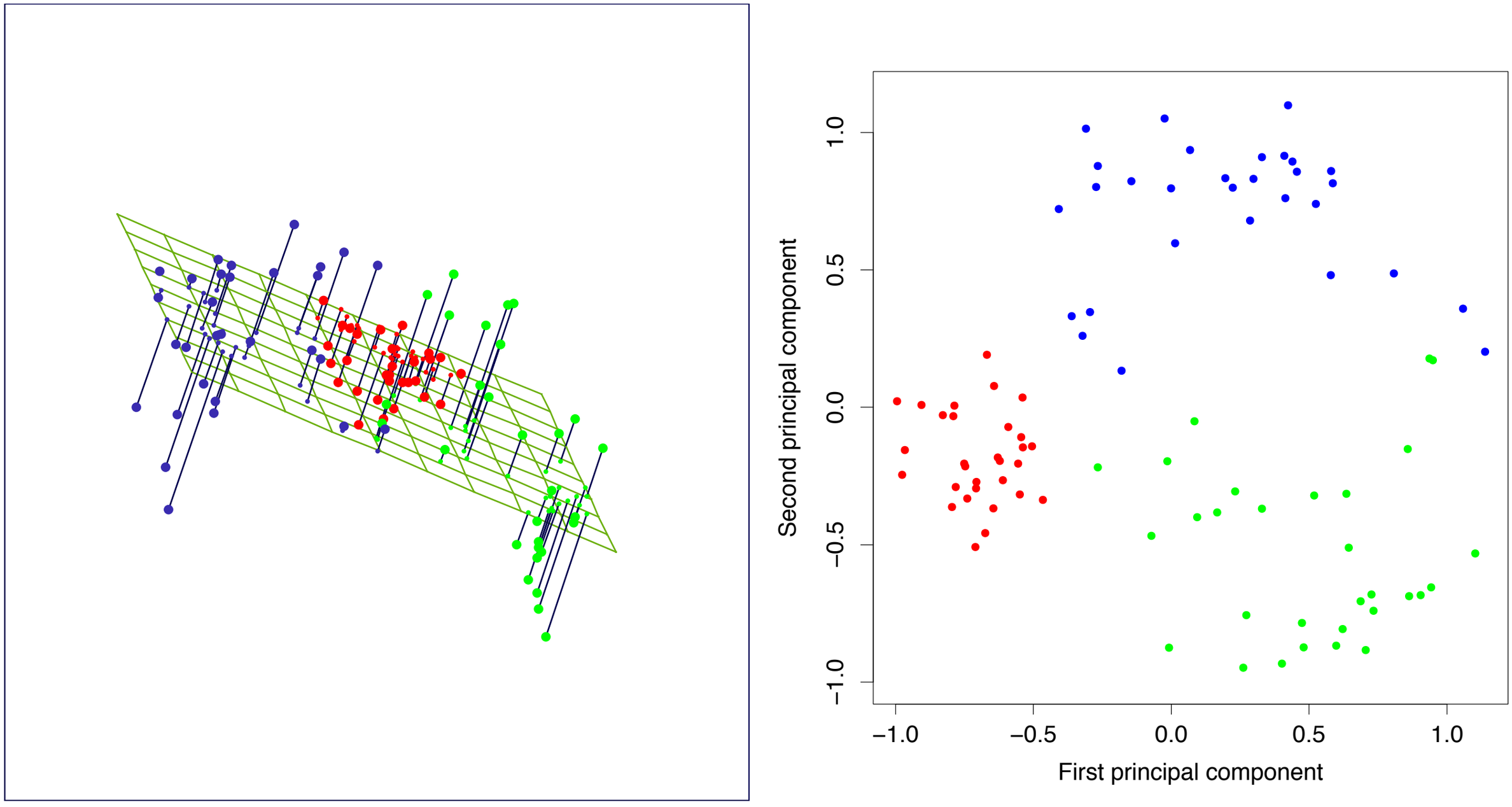


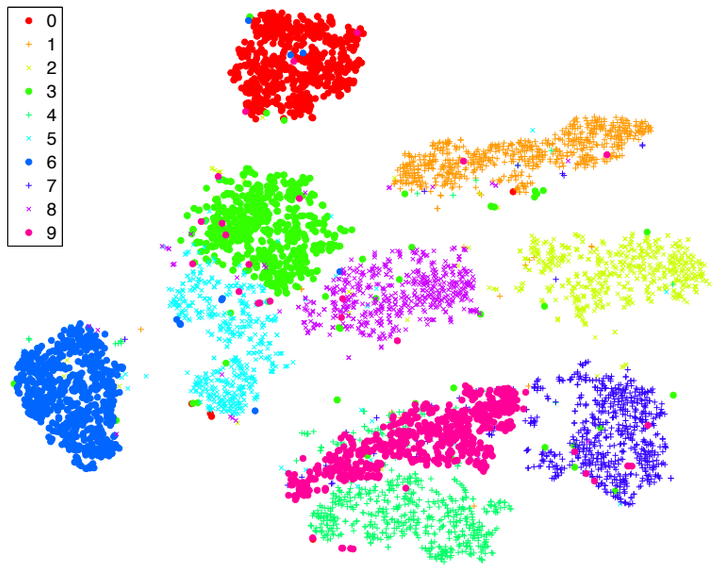
FIGURE 14.21. *The best rank-two linear approximation to the half-sphere data. The right panel shows the projected points with coordinates given by* $\mathbf{U}_2\mathbf{D}_2$*, the first two principal components of the data.*

Advanced topics

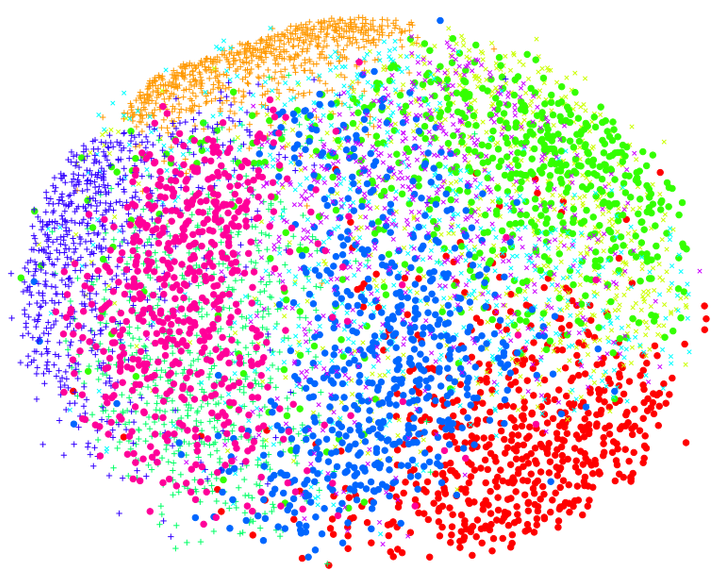
Multi-dimensional scaling: focus on (near) distances

Locally linear embedding (e.g. Donoho & Grimes, PNAS 2003)

t-SNE (van der Maaten & Hinton, JMLR 2008)

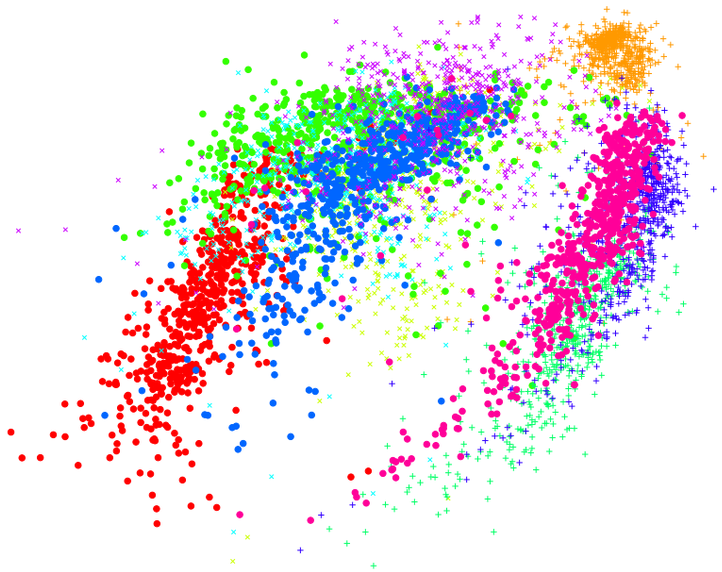


(a) Visualization by t-SNE.

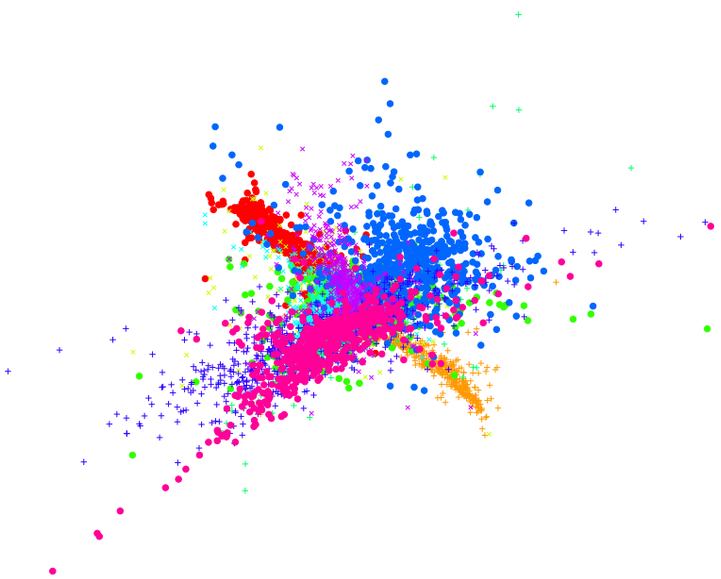


(b) Visualization by Sammon mapping.

Figure 2: Visualizations of 6,000 handwritten digits from the MNIST data set.

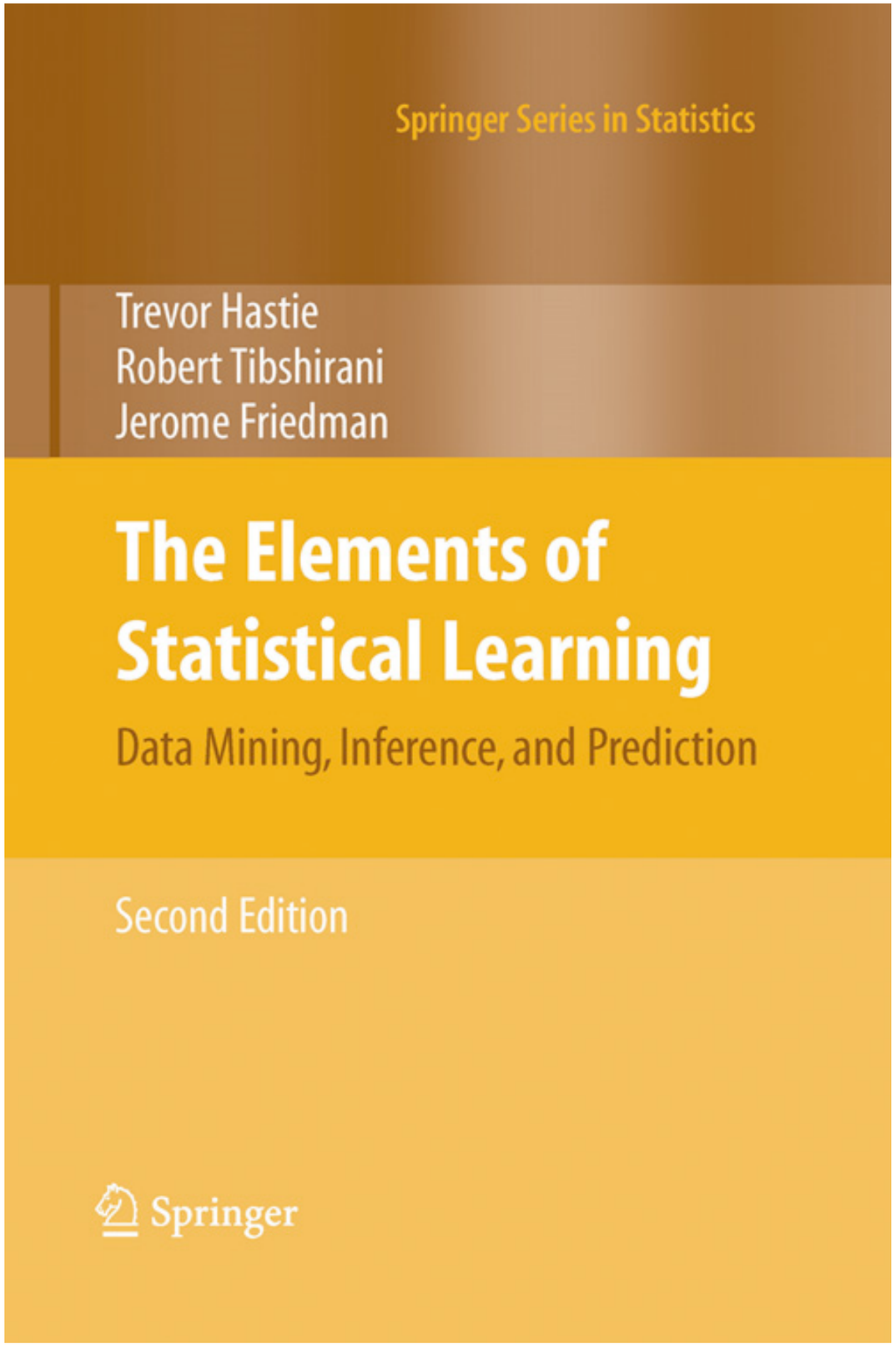


(a) Visualization by Isomap.



(b) Visualization by LLE.

Figure 3: Visualizations of 6,000 handwritten digits from the MNIST data set.



CRAN Task View: Analysis of Ecological and Environmental Data:
Ordination

http://cran.r-project.org/web/views/Environmetrics.html