Reduction

Mladen Victor WICKERHAUSER

Washington University in St. Louis, Missouri
victor@wustl.edu
http://www.math.wustl.edu/~victor

Dimensionality Reduction and Manifold Estimation
PMF — University of Zagreb
Winter, 2022

R Programming

Another useful language: R

- Public software, supported by thousands
- Main site: `http://www.r-project.org/`
- Integrated development environment (also free): `https://www.rstudio.com/products/rstudio/download/`

Data Visualization in R

Scatterplots, contours, and perspective.

- 2-D or 3-D with perspective
- Class website file `04datav.txt`
- Pairwise display of coordinates
- Class website file `04malda.txt`

Linear Model with Additive Error

Model:

$$y = \alpha + \beta_1 x_1 + \cdots + \beta_n x_n, \qquad \text{for large } n.$$

Data with errors:

$$y_i = \alpha + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i} + \eta_i, \quad i = 1, \ldots, m,$$

with $m \approx n$ and i.i.d. errors $\{\eta_i\}$.

Find best linear unbiased estimator $a, b_1, \ldots, b_n$ for $\alpha, \beta_1, \ldots, \beta_n$ by least-squares regression.

Stepwise Regression

Idea: eliminate irrelevant x variables.

- Multiple hypothesis test on

$$H_0 : \beta_1 = \cdots = \beta_n = 0$$

  If H_0 is rejected, then there is some dependence.
- Individual hypothesis tests of $H_0 : \beta_i = 0$.
- Class website `04stepr.txt`

Method: ANOVA and removal of x_i with insignificant β_i.

Principal Components and Linear Discriminants

Empirical Karhunen-Loeve on

$$\{(y_i, x_{1,i}, \ldots, x_{n,i}) : i = 1, \ldots, m)\} \subset \mathbf{R}^{n+1}$$

- Find the orthogonal directions of highest variance
- Supervised learning for multiple classes
- Class website file `04malda.txt`

Method: diagonalize the empirical covariance matrix.

Classification Trees

Example: gene expression data by cancer cells.

- Supervised learning
- Cross-Validation
- Class websitefile: `04trees.txt`

Clustering

Examples: Irises, cancers

- Unsupervised learning
- Agglomerative and Divisive clustering
- Class website file: `04clust.txt`

References

- Kim Seefeld and Ernst Linder. *Statistics Using R with Biological Examples*. (2007)