Streamwise Regression and CART

1 Sure Independence Screening

In this exercise we would like to get familiar with the R package *SIS* and employ this screening procedure presented in the course (see **paragraph 3.3.2**). After having read the documentation of the R package *SIS* and installed it, perform the following steps:

(a) Load the *Leukemia* dataset as explained in *Practical 2*.

(b) Split the dataset randomly in two halves (i.e. 36 observations each) and create a *train* and *test* sample.

(c) The functions SIS() performs first a screening procedure based on marginal correlations and then applies a penalized method (forthcoming *Chapter 4* of the e-book) to obtain the final model. Choose among all the available options (i.e. in terms both of penalized methods and tuning constants) three candidates and evaluate the predictions of the selected models on the *test* sample. Which penalized method performs best in this specific example after the SIS?

2 Programming the PC-simple algorithm

First of all we retrieve the simulation setting used in *Practical 5* with some changes:

1. Generate from a MVN (multivariate normal) a matrix $\mathbf{X_{n*p}}$ with $n = 1000$ and $p = 10$. You can choose the location vector as you wish but set the scale matrix with an autoregressive form $\mathbf{\Sigma} = [\sigma_{lm}]_{l,m=1,\dots,p}$ with $\sigma_{lm} = \rho^{|l-m|}$.
2. Fix $\rho = 0.5$ and set the seed equal to 11 (i.e. set.seed(11)).
3. Choose the generating vector $\boldsymbol{\beta} = [3\ 1.5\ 0\ 2\ rep(0, 6)]$.
4. Generate $\hat{\mathbf{y}}$ thanks to the relation $\mathbf{y} = \mathbf{X_{n*p}}\ \boldsymbol{\beta} + \boldsymbol{\epsilon}$ where ϵ_i is a standard normal. Suppose for simplicity that the errors are uncorrelated.

Now, after having read the PC-simple algorithm description in **paragraph 3.3.3**, perform the following passages on your simulated data:

(a) Find the active set M_1 using the Fisher's Z transformation and the associated correlation coefficient test (fix $\alpha = 0.05$ for the rest of the exercise).

(b) Calculate all the partial correlations of order 1 (i.e. one variable at the time) of the active set M_1, test them and retrieve the new active set $M_2 \subseteq M_1$.

(c) Calculate the partial correlations of higher order and test them until your reach the condition $M_{m-1} = M_m$ which implies the convergence of the PC-simple algorithm. Do you obtain the exact model?

3 Classification and Regression Trees

In this exercise we would like to get familiar with the R package *rpart*. After having read **paragraph 3.4** of the e-book, the documentation of the *rpart* package and having installed it, perform the following steps:

(a) Load the *Iris* dataset already present in R, split it randomly in a *train* and *test* sample (common choice is $\frac{2}{3}$ train and $\frac{1}{3}$ test).

(b) Fit a classification tree with the function rpart() and plot the tree. Have a look at *rpart.plot* package if you want to improve the appearance of the fitted tree.

(c) After having pruned the tree, evaluate its prediction on the *test* sample (i.e. use predict() on a tree object).

In order to build a regression tree, retrieve the *Zambia* dataset, consider only the continuous variables and performs the following steps:

(a) Load the *Zambia* dataset, split it randomly in a *train* and *test* sample (common choice is $\frac{2}{3}$ train and $\frac{1}{3}$ test).

(b) Fit a classification tree with the function rpart() and plot the tree. Have a look at *rpart.plot* package if you want to improve the appearance of the fitted tree.

(c) After having pruned the tree, evaluate its prediction on the *test* sample (i.e. use predict() on a tree object).