Exhaustive Search - CV and AIC

1 Simulations and CV

In this exercise we would like to program k-fold Cross-Validation (with k=2) and do model selection in a specific simulation setting with an exhaustive search.

(a) Generate from a MVN (multivariate normal) a matrix $\mathbf{X_{n*p}}$ with $n = 1000$ and $p = 5$. You can choose the location vector as you wish but set the scale matrix as the identity.

(b) Choose the generating vector $\boldsymbol{\beta} = [3\ 1.5\ 0\ 2\ 0]$ and retrieve the signal to noise ratio of this setting.

(c) Generate the responses 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.

(d) Split the data randomly in two halves (k=2) and use one half as a training set to determine $\hat{\boldsymbol{\beta}}_{\boldsymbol{MLE}}$. Then, making use of the specific loss function of the linear regression, find the test set cross validation error for each possible model with the other half. Repeat the procedure switching training and test set. Conclude on the best model by averaging the cross validation errors in the two data halves for each possible model.

(e) Suppose now that we increase the size of $\boldsymbol{\beta}$ to 100 (i.e. $p = 100$). Calculate the number of possible models together with an estimate of the time needed for an exhaustive search (*hint: use previous results*). Conclude on the feasibility of the task.

2 Simulations and AIC

In this exercise we would like to repeat the steps underlined above in the specific case of the Akaike information criterion.

(a) Retrieve the values found up to point (c) of exercise 1.

(b) Calculate the AIC for all possible models when $p = 5$ without using the predefined function present in R. Conclude on the best model.

(c) As for the previous exercise, suppose now that we increase the size of $\boldsymbol{\beta}$ to 100 (i.e. $p = 100$). Conclude on the feasibility of the task.

(d) Compare the performance of CV and AIC by replicating 100 times the tasks of Exercise 1 and 2. This implies generating a new vector $\mathbf{y} = \mathbf{X_{n*p}}\,\boldsymbol{\beta}+\boldsymbol{\epsilon}$ keeping fixed the matrix $\mathbf{X_{n*p}}$, the population vector $\boldsymbol{\beta}$ and the data random split for the 2-fold CV (i.e. only the ϵ_i are random in this setting). In particular you should compare CV and AIC in light of three specific criteria: the proportion of times the correct model is selected (*Exact*), the proportion of times the selected model contains the correct one (*Correct*) and the average number of selected regressors (*Average* $\sharp$).

3 Real Data Application (optional: does not count for the grade)

Load the *Zambia* dataset (see exercise 3 of Practical 1 for further info) and perform an exhaustive search on the continuous covariates (i.e. avoiding factors) based on CV and AIC in order to find the best model.

You can either employ your codes derived in Exercise 1 and 2 or make use of the existing R packages: **leaps**, **glmulti**, **MuMIn** and **caret**.